Titre: Title:	Correction par simulations de tests multiples dans les études d'association génomique familiale
Auteur: Author:	Pierre-Luc Brunelle
Date:	2008
Type:	Mémoire ou thèse / Dissertation or Thesis
Référence: Citation:	Brunelle, P.-L. (2008). Correction par simulations de tests multiples dans les études d'association génomique familiale [Mémoire de maîtrise, École Polytechnique de Montréal]. PolyPublie. https://publications.polymtl.ca/8325/

UNIVERSITÉ DE MONTRÉAL

CORRECTION PAR SIMULATIONS DE TESTS MULTIPLES DANS LES ÉTUDES D'ASSOCIATION GÉNOMIQUE FAMILIALE

PIERRE-LUC BRUNELLE
DÉPARTEMENT DE GÉNIE INFORMATIQUE ET DE GÉNIE LOGICIEL
ÉCOLE POLYTECHNIQUE DE MONTRÉAL

MÉMOIRE PRÉSENTÉ EN VUE DE L'OBTENTION
DU DIPLÔME DE MAÎTRISE ÈS SCIENCES APPLIQUÉES
(GÉNIE INFORMATIQUE)
JUIN 2008

Your file Votre référence
ISBN: 978-0-494-46039-9
Our file Notre référence
ISBN: 978-0-494-46039-9

UNIVERSITÉ DE MONTRÉAL

ÉCOLE POLYTECHNIQUE DE MONTRÉAL

Ce mémoire intitulé:

CORRECTION PAR SIMULATIONS DE TESTS MULTIPLES DANS LES ÉTUDES D'ASSOCIATION GÉNOMIQUE FAMILIALE

présenté par: BRUNELLE Pierre-Luc
en vue de l'obtention du diplôme de: Maîtrise ès sciences appliquées
a été dûment accepté par le jury d'examen constitué de:

M. ANTONIOL, Giuliano, Ph.D., président
M. MERLO, Ettore, Ph.D., membre et directeur de recherche
M. HAMET, Pavel, M.D., Ph.D., membre et codirecteur de recherche
M. ROCHELEAU, Ghislain, Ph.D., membre

Remerciements

Je tiens d'abord à remercier chaleureusement Ettore Merlo pour avoir dirigé mon projet de maîtrise. Il m'a offert à la fois une grande liberté dans le choix du sujet de recherche et de la méthode et une grande disponibilité aux moments importants. Je le remercie en particulier pour les conseils qu'il m'a offerts durant la rédaction du mémoire. La valeur qu'il accorde à son travail et à celui des autres m'a permis de mettre en relief les bons coups que j'ai réalisés. J'ai apprécié sa motivation, sa bonne humeur et sa rigueur scientifique.

Je tiens à remercier le Dr Pavel Hamet du Centre de recherche du Centre hospitalier de l'Université de Montréal pour avoir codirigé le projet. Il a assumé ma rémunération en tant qu'employé, financé le projet et mis à ma disposition toutes les données génotypiques et phénotypiques de la cohorte d'étude décrite dans ce mémoire. Sa grande curiosité et son intérêt à partager et à expliquer m'ont grandement aidé à mieux comprendre la génétique et la biologie.

Je remercie les membres du jury, Giuliano Antoniol, Ettore Merlo, Pavel Hamet et Ghislain Rocheleau, d'avoir pris le temps de réviser le mémoire.

Je veux remercier Bobby P. C. Koeleman, chercheur au Centre Universitaire Médical de Utrecht, Pays-Bas, pour m'avoir offert de l'encadrement et du support en statistiques. Ses interventions m'ont donné de l'assurance dans la méthode développée.

Je remercie Emmanuelle Guérin, ma conjointe, pour son support, son encouragement et sa grande confiance en moi.

Je remercie la compagnie Prognomix, mon employeur durant une partie de mes études, de m'avoir permis de me consacrer aux cours de maîtrise. Je remercie aussi les Instituts de recherche en santé du Canada (projet Cardiogène) pour le financement.

Je n'aurais pu réaliser les expériences décrites dans ce mémoire sans le travail effectué par les collaborateurs de Dr Hamet: les chercheurs Ondrej Seda et Johanne Tremblay ont

mis sur pied toute l'infrastructure de la base de données que j'ai utilisée. Aussi, j'ai bénéficié du génotypage qui a été fait sous la responsabilité du Dre Tremblay par Carole Long, Évelyne Morin et Gilles Corbeil.

Finalement, je remercie les membres du groupe de recherche du Dr Pavel Hamet pour leurs suggestions, commentaires et camaraderie: Audrey Noël, François Gauthier, Johanna Sandoval, Alexandru Gurau, Mahinè Ivanga, Ivan Arenas, Majid Nikpay et Louis-Philippe Lemieux-Perreault.

Résumé

L'équipe du Dr Pavel Hamet au CHUM étudie l'hypertension chez les familles du Saguenay-Lac-Saint-Jean. Une étude génomique a été lancée afin de déterminer les causes génétiques de l'hypertension. Plus de 58 000 polymorphismes d'un seul nucléotide (SNP) sont testés pour association chez 468 individus répartis en 76 familles. Un test statistique d'association génomique familiale est appliqué à chaque SNP.

Lorsqu'un seul test statistique est effectué, une valeur p inférieure ou égale à 0,05 est habituellement jugée statistiquement significative et permet de conclure que le polymorphisme est associé à la maladie. Par contre, dans le cas de tests multiples, l'utilisation d'un seuil de 0,05 résulte en un grand nombre de faux positifs. Par exemple, même si aucun des 58 000 polymorphismes étudiés n'a de relation avec l'hypertension, on peut s'attendre à ce qu'environ 5% d'entre eux, soit 2 900, rapportent une valeur p inférieure ou égale à 0,05. Il est nécessaire d'appliquer une correction pour tests multiples afin de contrôler le taux d'erreur au sein de la famille de tests.

La correction de Bonferroni est couramment employée mais s'avère conservatrice car elle ne tient pas compte de la corrélation entre les tests. La principale source de corrélation provient du déséquilibre de liaison (LD) entre les polymorphismes rapprochés. Les méthodes de rééchantillonnage, quant à elles, tiennent compte du LD, ce qui les rend plus puissantes. Par contre, les méthodes de rééchantillonnage présentement disponibles ne s'appliquent pas aux familles étudiées: certaines des familles sont très grosses (près de 100 individus) et plusieurs parents ne sont pas génotypés. Il est donc nécessaire de développer une correction pour tests multiples d'association génomique familiale qui soit puissante et s'applique aux familles étudiées.

Nons avons conçu et implanté une méthode novatrice basée sur le rééchantillonnage. Nous rééchantillonnons les génotypes sans modifier les phénotypes. La difficulté réside dans le fait que les génotypes générés doivent respecter la transmission mendélienne,

c'est-à-dire qu'un individu a reçu un allèle de son père et un allèle de sa mère à chaque polymorphisme autosomique. Nous surmontons cette difficulté en simulant pour les fondateurs des génotypes qui respectent les caractéristiques des génotypes originaux puis en transmettant ces génotypes à leurs descendants par *gene-dropping*.

Notre méthode est très générale: les familles peuvent être de taille arbitraire; tous les patrons de génotypes manquants sont supportés, en particulier la situation dans laquelle les parents ne sont pas génotypés; aucune variabilité phénotypique n'est nécessaire; la correction peut être appliquée à tout test d'association génomique familiale; et des dizaines de milliers de SNPs peuvent être traités. À notre connaissance, nous proposons la première méthode empirique de correction pour tests multiples dans les études d'association génomique familiale qui supporte des données aussi générales.

Nous avons développé un logiciel qui implante la méthode proposée. Quatre critères ont guidé le développement du logiciel: la rapidité de développement, la rapidité de l'exécution des simulations, l'aisance de gestion des expériences et des résultats et la fiabilité. Ainsi, nous avons employé un langage de programmation de haut niveau, Python, intégré et validé des composantes externes, implanté un calcul distribué qui permet une accélération presque linéaire et conçu et intégré une base de données relationnelle qui stocke les expériences et les résultats. Les principales activités de génie logiciel que nous avons menées sont: la conception du logiciel, son implantation et sa validation. La validation montre que la méthode proposée contrôle adéquatement le taux d'erreur au sein de la famille de tests.

Huit expériences ont été effectuées sur différents phénotypes reliés à l'hypertension et différentes régions autosomiques. La méthode proposée s'avère moins sévère que la correction de Bonferroni, jusqu'à 30 fois dans certains cas. Nous avons découvert une association statistiquement significative qui aurait été manquée par la correction de Bonferroni: le SNP rs10494966 est associé au pourcentage de gras corporel calculé à partir des plis cutanés ($p_{\text{corrigée}} = 0{,}037$). Quatre vérifications ont montré que l'association n'est pas due à une erreur de génotypage.

Abstract

Dr Pavel Hamet and his team at CHUM are studying hypertension in French-Canadian families from Saguenay-Lac-Saint-Jean. A genomewide association study was launched to find the genetic polymorphisms that determine hypertension and related phenotypes. The study investigated over 58,000 single nucleotide polymorphisms (SNPs) genotyped in 468 subjects from 76 families. A statistical test has been performed for each SNP.

Usually, when a single statistical test is performed, a p-value lower than 0.05 is interpreted as proof that the SNP is associated with the disease. However, this cutoff is inappropriate when performing tests on many SNPs because many false positives would be reported. Even if none of the SNPs were associated to hypertension, about 5% (i.e. 2,900), would yield a p-value below 0.05. Thus it is necessary to apply some multiple-testing correction to control the family-wise error rate (FWER).

The Bonferroni correction is one of the easiest multiple-testing corrections available. Unfortunately, the Bonferroni correction fails to take into account the correlation between tests and therefore it is excessively conservative. The correlation among tests mainly stems from the linkage disequilibrium (LD) among nearby SNPs. Resampling methods are able to take this correlation into account. However, currently available resampling methods cannot handle the families under study. Some families are large (nearly 100 subjects) and many parents have not been genotyped. Our goal is to develop a powerful multiple testing correction method that can handle such families.

The main idea is to resample the genotypes while keeeping the phenotypes fixed. We assign to each individual genotypes that match the original genotypes properties such as the LD. There is one problem though: the resampled genotypes must obey Mendel genetic rules, which means that one individual received one allele from the father and one from the mother. We solve this problem by assigning to founders genotypes that match the original genotypes properties and then transmitting those genotypes by gene-

dropping.

The method we have developed is very general. It handles large families, all missing genotype patterns, including the pattern in which parents have not been genotyped, phenotypes without any variability, any underlying family-based association test and large numbers of SNPs (tens of thousands) in a reasonable amount of time. To the best of our knowledge, we are proposing the first empirical multiple testing correction method for family-based association tests that can handle such general data.

We built a software to implement the method. We tried to respect four design goals: the software development had to be fast, the simulations had to be reasonably fast, the experiments had to be easily stored and retrieved and the results had to be reliable. To achieve those design goals, we used a high-level programming language, Python, integrated and validated third-party software and libraries, implemented the software so the computation could be distributed on several computers to achieve near-linear speedup and developed a relational database to store the experiments and results. The main software engineering activities that we carried were design, implementation and validation. The validation showed that our method correctly controls the FWER.

We performed eight experiments between phenotypes related to hypertension and many autosomal regions. Results show that our method can be up to 30 times less severe than the Bonferroni correction. Furthermore, our method detected a statistically significant association after multiple testing correction. The association would have been missed by Bonferroni correction. SNP rs10494966 is associated to body fat by skinfolds ($p_{corrected} = 0.037$). We verified that the association was not caused by a genotyping error.

Table des matières

Liste des tableaux

Liste des figures

Liste des sigles, abréviations et unités de mesures

cM	centiMorgans
CNV	*copy number variation*, variabilité du nombre de copies
FDR	*false discovery rate*, taux de fausses découvertes
pFDR	*positive false discovery rate*, taux positif de fausses découvertes
FPRP	*false positive report probability*, probabilité qu'un résultat statistiquement significatif soit faux
FWER	*family-wise error rate*, taux d'erreur au sein de la famille de tests
Go	gigaoctet (2^{30} octets)
GWAS	*genome-wide association study*, étude d'association génétique à large échelle
HMM	*hidden Markov model*, modèle de Markov caché
HWE	*Hardy-Weinberg equilibrium*, équilibre de Hardy-Weinberg
kb	mille paires de bases
LCG	*linear congruential generator*, générateur congruentiel linéaire
LD	*linkage disequilibrium*, déséquilibre de liaison
Mb	un million de paires de bases
pb	paire de bases
PRNG	*pseudo-random number generator*, générateur de nombres pseudo-aléatoires
RAM	*random access memory*, mémoire vive
SNP	*single nucleotide polymorphism*, polymorphisme d'un nucléotide

Chapitre 1: Introduction

Depuis quelques années, de nombreuses études d'association génétiques sont publiées. Ces études tentent de démontrer l'association entre un marqueur génétique et une maladie ou, plus généralement, un phénotype, permettant ainsi d'identifier les facteurs génétiques qui sont responsables du phénotype. Ces études ont été rendues possibles grâce à l'apparition de puces de génotypage de plus en plus denses. Les puces les plus récentes déterminent le génotype d'un individu à environ un million de marqueurs génétiques.

Pour déterminer une association, un test statistique est effectué entre les génotypes des individus et leurs phénotypes. Au moins un test est effectué pour chaque marqueur génétique. Une étude peut examiner plusieurs phénotypes, examiner différents sous-groupes (hommes, femmes, obèses, non-obèses, etc.) et utiliser plusieurs tests statistiques et plusieurs modèles génétiques. Ainsi, il n'est pas rare que des centaines de milliers, voire des millions de tests statistiques, soient effectués. Nous sommes confrontés à un problème de tests multiples. Traditionnellement, un seuil de significativité de 0,05 est utilisé pour déterminer qu'une association est statistiquement significative. Ainsi, une valeur p inférieure ou égale au seuil amène à rejeter l'hypothèse nulle (pas d'association) et à accepter l'hypothèse alternative (association). En supposant qu'un million de marqueurs génétiques sont testés mais que seulement quelques uns sont associés au phénotype, environ 5% rapporteront une valeur p inférieure ou égale à 0,05: 50 000 associations « significatives » seront rapportées, la plupart faussement!

Pour éviter de rapporter autant de résultats faussement positifs, le seuil de significativité doit être modifié. Plusieurs types d'erreur peuvent être contrôlés. Le taux d'erreur au sein de la famille de tests (*family-wise error rate*, FWER) représente la probabilité de rejeter au moins une hypothèse nulle vraie. Plusieurs méthodes existent pour contrôler le FWER. La correction de Bonferroni est couramment employée et simple à utiliser: il suffit de diviser le FWER par le nombre de tests effectués et utiliser ce seuil pour

chaque test. Par exemple, si le FWER souhaité est de 0,05 et que cinq tests sont effectués, un test devra produire une valeur p inférieure ou égale à 0,01 afin d'être déclaré significatif.

Malheureusement, la correction de Bonferroni s'avère conservatrice lorsque les tests sont corrélés. En d'autres termes, il est plus difficile de détecter une vraie association. La corrélation dans les études génétiques provient de plusieurs sources. D'abord, les marqueurs génétiques qui sont rapprochés peuvent représenter la même information, ce qui est appelé déséquilibre de liaison (*linkage disequilibrium*, LD). Ensuite, des phénotypes peuvent être corrélés entre eux, par exemple le poids et l'indice de masse corporelle. Troisièmement, les modèles génétiques utilisés ne sont pas indépendants les uns des autres. Finalement, différents sous-groupes qui ne sont pas mutuellement exclusifs sont parfois étudiés.

Pour augmenter les chances de détecter une vraie association, différentes méthodes de simulation ont été développées. Elles incorporent les caractéristiques des données originales, en particulier les corrélations. Pour les études cas-témoins, une solution simple est de permuter le statut d'affection entre tous les individus sans modifier les génotypes; la procédure est répétée des milliers de fois. Cette approche conserve la corrélation entre les marqueurs génétiques. Des méthodes ont également été développées pour les études de trios et les études familiales simples. Par contre, aucune méthode n'est satisfaisante dans le cas de grandes familles avec des patrons de génotypes manquants arbitraires.

Le groupe de recherche du Dr Pavel Hamet établi au Centre de recherche du Centre hospitalier de l'Université de Montréal (CR-CHUM) s'intéresse à l'hypertension dans la population du Saguenay-Lac-Saint-Jean. Près de 900 sujets, répartis dans une centaine de familles, ont été recrutés, phénotypés et environ la moitié ont été génotypés. Certaines de ces familles sont très grandes (parfois près de 100 sujets). Présentement, aucune méthode de correction pour tests multiples basée sur les simulations ne peut traiter adéquatement d'aussi grandes familles. Le but de ce mémoire est de présenter une

méthode novatrice de correction pour tests multiples qui s'applique à ces familles.

Le problème n'est pas simple. On peut soit permuter les phénotypes et ne pas modifier les génotypes, ou encore permuter les génotypes et ne pas modifier les phénotypes. Dans les deux cas, une permutation naïve ne serait pas valide: si on permute librement les phénotypes, on perd l'héritabilité (c'est-à-dire la corrélation phénotypique chez les membres d'une même famille) alors que si on permute librement les génotypes, on introduit des erreurs de transmission mendélienne.

La méthode que nous proposons repose sur l'idée de générer pour chaque individu des génotypes qui respectent les caractéristiques des génotypes originaux et de ne pas modifier les phénotypes. Afin de respecter la transmission mendélienne et de conserver le déséquilibre de liaison, nous simulons pour les fondateurs des génotypes qui respectent ce déséquilibre de liaison puis transmettons leurs génotypes à leurs descendants par la méthode du *gene-dropping*. La corrélation entre les phénotypes est également maintenue puisqu'ils ne sont pas permutés.

Nous avons développé un logiciel qui implante la méthode proposée. Quatre critères ont guidé le développement du logiciel. Premièrement, le développement doit être rapide: ainsi, si la méthode s'avère inadéquate, on pourra proposer une autre méthode. Deuxièmement, l'exécution du logiciel doit être assez rapide afin d'obtenir des valeurs p empiriques précises dans un délai raisonnable. Troisièmement, les expériences doivent être repérables et on doit pouvoir facilement comparer les résultats entre expériences et entre méthodes de correction pour tests multiples. Finalement, le logiciel doit être fiable afin que les chercheurs qui l'utilisent aient confiance dans les résultats rapportés.

Huit expériences sur des puces de génotypage Xba de la compagnie Affymetrix, qui rapportent le génotype de 58 000 marqueurs, et sur divers phénotypes sont effectuées afin de comparer la méthode proposée à la correction de Bonferroni et déterminer si la méthode proposée peut rapporter des associations statistiquement significatives après correction pour tests multiples. La méthode peut être appliquée à tout test statistique d'association génomique familiale. Nous avons choisi le test FBAT (*family-based*

association test) pour les besoins de nos expériences.

1.1 Définition du problème

Nous donnons dans cette section une définition sommaire du problème que nous voulons résoudre. Nous repoussons à la section 4.4 la définition formelle du problème, après que certaines notions génétiques et statistiques aient été présentées.

Nous voulons développer une méthode de correction pour tests multiples qui s'applique aux études d'association génomique familiale. Nous voulons contrôler le nombre de fausses associations rapportées, tout en essayant de rapporter le plus de vraies associations. Tous les tests statistiques d'association familiale doivent être supportés.

1.2 Division du mémoire

Ce mémoire se divise ainsi. Aux chapitres 2 et 3, nous présentons les concepts génétiques nécessaires à la compréhension de la méthode proposée et du test d'association génomique familiale employé. Au chapitre 4, nous rappelons certaines notions statistiques, décrivons notre problème de façon détaillée puis présentons les méthodes de correction multiples présentement employées et expliquons en quoi chacune est inadéquate. Au chapitre 5, nous présentons la méthode que nous avons inventée pour corriger par simulations les tests multiples d'association génomique familiale. Nous décrivons au chapitre 6 les critères qui ont guidé le développement de notre logiciel, les principales activités de génie logiciel que nous avons menées à bien et les choix de conception et d'implantation que nous avons faits. Nous présentons au chapitre 7 les données sur lesquelles nous travaillons, les expériences que nous avons effectuées et les résultats que nous avons obtenus. Ces expériences visent deux buts: trouver des associations statistiquement significatives et comparer notre méthode à la correction de Bonferroni. Nous terminons ce mémoire par une discussion (chapitre 8) et par des conclusions et pistes de recherche futures (chapitre 9).

Chapitre 2: Notions génétiques

Dans ce chapitre, nous faisons un rappel des notions génétiques indispensables à la compréhension de la méthode. La référence principale de ce chapitre est Ott (1991).

2.1 Gènes et chromosomes

Le bagage génétique chez les êtres humains se répartit en 23 paires de chromosomes: 22 autosomes ainsi que les chromosomes sexuels X et Y. Un individu a reçu un jeu de chromosomes de son père et un de sa mère. Les deux chromosomes d'une paire sont appelés des chromosomes homologues. Les caractères héréditaires, tels la taille et la pression artérielle, sont en partie déterminés par des gènes; ces caractères sont appelés des phénotypes. Les être humains possèdent entre 20 000 et 25 000 gènes selon des estimations récentes (International Human Genome Sequencing Consortium, 2004). Sauf dans le cas d'aberrations génétiques, deux individus possèdent les mêmes gènes; par contre, la forme d'un gène peut être différente chez les deux individus. Les différentes formes d'un gène sont appelées allèles. La paire d'allèles d'un individu à un gène constitue son génotype. Un individu possédant deux allèles identiques à un gène est dit homozygote pour ce gène, alors qu'un individu possédant deux allèles différents est dit hétérozygote. Le terme locus indique un endroit quelconque sur un chromosome, que ce soit à l'intérieur d'un gène ou non.

La méiose est le processus menant à la création de cellules sexuelles (gamètes) qui contiennent une seule copie de chaque chromosome à partir de cellules possédant deux copies de chaque chromosome. À cette étape, deux phénomènes augmentent la diversité génétique d'une population. D'abord, les 23 paires de chromosomes d'une cellule mère s'apparient de façon indépendante dans les cellules filles. Il y a 2^{23} appariements possibles. Ensuite, l'enjambement (*crossover*) échange une partie du chromosome paternel contre une partie équivalente du chromosome maternel homologue. Il semble toujours y avoir au moins un enjambement par chromosome par méiose (Ott, 1991, p.

13). Selon le terme de James Watson, les enjambements se manifestent de façon semi-aléatoire: la distribution des enjambements n'est pas uniforme le long d'un chromosome (au contraire, il y a des points chauds et des points froids) et la manifestation d'un enjambement semble supprimer la manifestation d'autres enjambements dans la même région.

Les allèles à plusieurs loci reçus d'un parent constituent un haplotype (Ott, 1991, p. 5). Si ces allèles proviennent tous d'un seul grand-parent, l'haplotype est non-recombinant, alors que si les allèles proviennent des deux grands-parents, l'haplotype est recombinant; on parle alors de recombinaison entre les loci. L'appariement indépendant des chromosomes cause la recombinaison entre les loci situés sur différents chromosomes, alors que l'enjambement cause la recombinaison entre les loci d'un même chromosome. Plus deux loci sont rapprochés, plus la probabilité qu'un enjambement survienne entre eux est faible. La distance génétique, exprimée en Morgans (M), est le nombre attendu d'enjambements entre deux loci pour un brin de chromosome. Cette distance est plus couramment exprimée en centiMorgans (cM). Or, il n'est pas possible d'observer directement les enjambements; par contre, il est possible d'observer les recombinaisons. La fraction de recombinaison, notée θ, représente la probabilité qu'un gamète produit par un parent soit un recombinant. Le nombre prévu d'enjambements et la fraction de recombinaison ne sont pas nécessairement identiques: un nombre impair d'enjambements cause une recombinaison mais un nombre pair d'enjambements s'annule. Plus deux loci sont éloignés, plus la probabilité que deux ou plusieurs enjambements surviennent entre eux augmente et plus la fraction de recombinaison augmente. La relation entre la distance génétique et la fraction de recombinaison est fournie par une fonction de carte (*map function*). Deux fonctions couramment employées sont celles proposées par Haldane en 1919 et par Kosambi en 1944. Dans ces fonctions, la fraction de recombinaison augmente de façon logarithmique par rapport au nombre prévu d'enjambements et tend vers ½ lorsque le nombre prévu d'enjambements tend vers l'infini.

Deux loci pour lesquels la fraction de recombinaison est inférieure à ½ sont dits liés.

L'analyse de liaison a pour but de déterminer si deux loci sont liés et, dans l'affirmative, de déterminer leur fréquence de recombinaison.

L'analyse d'association, quant à elle, vise à déterminer si un marqueur génétique est associé au phénotype, c'est-à-dire si le marqueur est directement responsable du phénotype ou proche d'un gène qui détermine le phénotype (Lazzeroni & Lange, 1998, p. 69). Ce type d'analyse est basé sur le déséquilibre de liaison, décrit plus loin.

2.2 Marqueurs génétiques

Les 23 chromosomes sont constitués d'une longue chaîne de paires de nucléotides en forme de double hélice. Le chromosome le plus long contient environ 240 millions de paires de nucléotides. L'ensemble des chromosomes en contient près de trois milliards, dont seulement 34 millions, soit 1%, se trouvent dans des gènes (International Human Genome Sequencing Consortium, 2004). Les variations génétiques ne se produisent pas uniquement dans les gènes. Un marqueur génétique est un locus polymorphe, c'est-à-dire pour lequel il existe des différences entre les individus. De façon analogue aux gènes, nous parlerons d'allèles et de génotypes à ces marqueurs. Plusieurs types de marqueurs ont été découverts puis utilisés dans les études de liaison et d'association. Les polymorphismes d'un seul nucléotide (*single nucleotide polymorphisms*, SNPs) sont présents dans la population sous deux formes. On estime à 10 millions le nombre de SNPs relativement fréquents chez l'être humain[1]. Leur grand nombre en font des marqueurs de choix pour les études d'association génomiques. Il y a quatre types de nucléotides: adénine (A), cytosine (C), guanine (G), et thymine (T). Quant à eux, les marqueurs microsatellites, aussi connus sous le nom de polymorphismes associés aux séquences répétées en tandem (*variable number of tandem repeats*, VNTR), sont davantage polymorphes que les SNPs, ce qui les rend appropriés aux études de liaison.

1 hapmap.org/whatishapmap.html

2.2.1 Conséquences des SNPs sur certaines maladies

L'approche d'analyse d'association par examen de tout le génome (*genome-wide association studies*, GWAS) a récemment fourni des résultats statistiquement très significatifs pour plusieurs maladies. Par exemple, Sladek et al. (2007) ont rapporté des associations entre le diabète de type II et huit SNPs provenant de cinq régions. Les valeurs p non-corrigées varient de 10^{-34} à 10^{-4}. Également pour le diabète de type II, Diabetes Genetics Initiative et al. (2007) ont rapporté neuf SNPs significatifs dans huit régions; les valeurs p varient de 10^{-48} à 10^{-6}. Certaines des associations trouvées par une étude ont été répliquées par l'autre étude. Maller et al. (2006) ont mis à jour des associations entre trois gènes et la dégénérescence maculaire liée à l'âge (DMLA), qui expliquent environ la moitié du risque de contracter la maladie. Ces trois études ont en commun d'avoir utilisé des puces de génotypage comprenant de trois cent mille à cinq cent mille SNPs et d'être de type cas-témoins, c'est-à-dire que les sujets de l'étude ne sont pas de proches parents les uns des autres.

2.3 Déséquilibre de liaison

Des marqueurs sont en déséquilibre de liaison (*linkage disequilibrium*, LD) lorsque les fréquences de leurs haplotypes ne sont pas égales à la multiplication des fréquences alléliques: les allèles de différents marqueurs ne sont pas indépendants les uns des autres (Ott, 1991, p. 5). Le LD a deux conséquences sur les études d'association: il permet de détecter un polymorphisme causant la maladie même si ce polymorphisme n'est pas directement observé et il induit une corrélation entre les marqueurs rapprochés.

Le LD est essentiel au succès des études d'association de tout le génome. En général, le polymorphisme qui cause la maladie n'est pas directement génotypé. Par contre, avec l'avènement de puces de génotypage de plus en plus denses, les chances sont élevées qu'au moins un des marqueurs de la puce se trouve à proximité du polymorphisme qui cause la maladie, donc que les deux loci soient en LD. En testant pour l'association entre le marqueur et la maladie, c'est l'association entre le polymorphisme qui cause la

maladie et la maladie qui est indirectement testée. Plus grand est le LD entre le marqueur et le polymorphisme, plus grande sera la capacité de détecter l'association.

La seconde conséquence du LD est que les tests d'association effectués sur des marqueurs rapprochés sont corrélés. Dans le cas d'un déséquilibre de liaison parfait entre deux marqueurs (c'est-à-dire que la connaissance de l'allèle à un marqueur permet de déterminer exactement l'allèle à l'autre marqueur), les résultats d'association aux deux marqueurs sont identiques; ce ne sont pas deux tests statistiques distincts qui sont effectués mais bien deux fois le même test. Une méthode de correction pour tests multiples qui ne tient pas compte de cette corrélation est conservatrice.

Au moins deux mesures du LD ont été proposées: D' et r^2 (Devlin & Risch, 1995). Considérons deux marqueurs, A et B, chacun possédant deux allèles, A_1 et A_2 pour le marqueur A et B_1 et B_2 pour le marqueur B. Notons p_1 et p_2 les fréquences des allèles A_1 et A_2 et q_1 et q_2 les fréquences des allèles B_1 et B_2. La fréquence attendue de l'haplotype A_1B_1, en absence de déséquilibre de liaison, est p_1q_1. Notons sa fréquence observée par x_{11}. Lewontin a modélisé en 1964 la différence entre la fréquence observée et la fréquence attendue:

$$D = x_{11} - p_1 q_1 \tag{2.1}$$

Puisque la valeur de D dépend des fréquences alléliques, on peut la normaliser pour obtenir une variable qui varie entre 0 (équilibre de liaison parfait) et 1 (déséquilibre de liaison parfait):

$$D' = \frac{D}{D_{max}} \tag{2.2}$$

dans lequel D_{max} représente la valeur maximale que D peut prendre étant donné les fréquences alléliques observées:

$$D_{max} = \begin{cases} min(p_1 q_2,\ p_2 q_1) & si\ \ D>0 \\ min(p_1 q_1,\ p_2 q_2) & si\ \ D<0 \end{cases} \tag{2.3}$$

L'autre mesure de LD est r^2, proposée par Hill et Robertson en 1968:

$$r^2 = \frac{D^2}{p_1 p_2 q_1 q_2} \tag{2.4}$$

Puisque les technologies actuellement utilisées ne permettent pas d'observer directement les haplotypes (elles rapportent seulement le génotype à chaque marqueur), il est nécessaire de phaser les génotypes, c'est-à-dire d'appliquer une procédure pour obtenir deux haplotypes à partir d'une séquence de génotypes.

2.4 Équilibre de Hardy-Weinberg

Les technologies actuelles commettent des erreurs dans les génotypes qu'elles rapportent, de l'ordre d'environ 1%. Un test permettant de détecter les erreurs de génotypage massives est le test de Hardy-Weinberg, qui compare les fréquences génotypiques observées aux fréquences attendues sous certaines conditions.

Lorsque certaines conditions sont respectées, principalement lorsque l'accouplement entre les individus d'une population est aléatoire et qu'il n'y a pas de forces extérieures telles la migration, la mutation ou une pression sélective au marqueur étudié, un marqueur est dit être en équilibre de Hardy-Weinberg, c'est-à-dire que les fréquences génotypiques ne dépendent que des fréquences alléliques (Ott, 1991, p. 3). Dans le cas de SNPs dont les allèles sont notés A et B et dont les fréquences alléliques sont notées p_A et p_B, les fréquences des génotypes AA, AB et BB sont p_A^2, $2p_Ap_B$ et p_B^2, respectivement.

Si les fréquences génotypiques observées à un marqueur diffèrent des fréquences génotypiques attendues sous l'équilibre de Hardy-Weinberg, cela peut vouloir dire que l'accouplement n'est pas aléatoire, que des forces extérieures s'appliquent, ou encore que le marqueur n'a pas été correctement génotypé. Dans le dernier cas, la raison peut être que le marqueur se situe dans un locus avec variabilité du nombre de copies (CNV; Redon et al., 2006), qu'une erreur de manipulation au laboratoire a eu lieu ou que le logiciel chargé de déterminer les génotypes a commis une erreur (Affymetrix, 2006). À

l'inverse, le fait que le marqueur est en équilibre de Hardy-Weinberg augmente notre confiance que les génotypes observés représentent bien les génotypes réels des individus.

2.5 Remarques sur l'indépendance

Notons que les concepts de liaison, de déséquilibre de liaison et d'équilibre de Hardy-Weinberg traitent tous les trois d'indépendance: entre marqueurs (liaison), entre allèles de différents marqueurs (LD), ou entre allèles d'un même marqueur (équilibre de Hardy-Weinberg). Dans les trois cas, il est possible d'appliquer un test statistique qui compare la valeur observée à la valeur attendue et de déterminer si une telle différence est fréquente lorsque l'hypothèse nulle d'indépendance est vraie.

Chapitre 3: Test d'association FBAT

Les études d'association se divisent en deux grandes catégories: les études cas-témoins, qui recrutent des individus faiblement apparentés, et les études familiales, qui recrutent plusieurs familles. Ces types d'études n'utilisent pas les mêmes tests statistiques pour découvrir des associations marqueur-phénotype. Les études cas-témoins font appel à des tests statistiques conventionnels, par exemple le test T de Student ou le test de chi-carré. Ces tests ne sont pas appropriés dans les études familiales car les observations (les génotypes et les phénotypes des individus) ne sont pas indépendantes, contrairement à ce que ces tests supposent (Newman et al., 2001). On utilise plutôt des tests qui comparent les génotypes observés aux génotypes attendus; les génotypes attendus sont calculés en faisant l'hypothèse que le marqueur étudié n'est pas associé au phénotype.

Les données sur lesquelles nous appliquons notre méthode, décrites en détails à la section 7.1, proviennent d'une étude familiale. L'association entre les SNPs et les phénotypes est testée au moyen du test TDT (*transmission disequilibrium test*), implanté dans le logiciel *FBAT* (*family-based association test*; Laird, 2006). *FBAT* supporte les phénotypes continus et binaires. Dans le cas des phénotypes binaires, les sujets non-affectés sont pris en compte dans l'analyse, une extension par rapport au test TDT original. Les génotypes des parents peuvent être inférés s'ils sont manquants. Les familles de toutes tailles sont supportées.

L'idée de base du test TDT est de comparer, chez les enfants malades, les génotypes observés aux génotypes attendus, ceux-ci étant calculés sous l'hypothèse qu'il n'y a aucune association entre la maladie et le marqueur. Si on s'aperçoit qu'un allèle est plus souvent transmis à des enfants malades que ce à quoi on s'attend, on peut conclure que le marqueur est associé à la maladie.

3.1 Statistique de test FBAT

La statistique de test FBAT est basée sur les génotypes des parents et des enfants et les phénotypes des enfants. Elle peut être calculée pour chaque allèle séparément ou pour tous les allèles simultanément. Nous présentons le développement relatif au test séparé de chaque allèle.

Soit T_{ij} une transformation de la valeur phénotypique Y_{ij} du sujet j de la famille i et X_{ij} une transformation de son génotype. T_{ij} et X_{ij} prennent des valeurs réelles. La statistique de test FBAT est définie comme étant (Laird, 2006):

$$U = S - E[S] \tag{3.1}$$

dans lequel:

$$S = \sum_{ij} T_{ij} X_{ij} \tag{3.2}$$

L'espérance de S, notée $E[S]$ est calculée sous l'hypothèse nulle (i.e. aucune association) et de façon conditionnelle par rapport aux génotypes parentaux.

Il est possible de calculer V, la variance de S. La statistique de test

$$Z = U/\sqrt{V} \tag{3.3}$$

est approximativement distribuée selon une loi normale de moyenne 0 et de variance 1. On peut donc comparer la valeur Z obtenue à la distribution normale pour savoir si une telle valeur Z s'observe fréquemment lorsqu'il n'y a pas d'association entre la maladie et le SNP. Il faut toutefois être prudent: la statistique de test est *approximativement* distribuée selon une loi normale; l'approximation n'est valide que lorsque la taille de l'échantillon est assez grande. Laird (2006) recommande de ne tester que les marqueurs pour lesquels au moins 10 familles sont informatives.

3.2 Codification du phénotype

La codification du phénotype a une grande influence sur la puissance du test. Pour que

le test soit valide, la codification de T_{ij} ne doit pas dépendre des génotypes. Laird (2006) offre plusieurs suggestions pour maximiser la puissance du test:

- Ajuster pour les covariables qui ont un effet sur le phénotype. Pour la plupart des phénotypes que nous étudions, nous ajustons pour l'âge et le sexe.
- Si le phénotype est binaire, soustraire à Y_{ij} la prévalence de la maladie dans la population, ou encore la prévalence dans l'échantillon. De manière analogue, si le phénotype est continu, soustraire la moyenne phénotypique de la population ou de l'échantillon. Nous nous sommes servis des prévalences et moyennes dans notre échantillon.

3.3 Modèles génétiques

La codification du génotype dépend du modèle génétique supposé par l'investigateur. *FBAT* permet de spécifier quatre modèles: additif, dominant, récessif et génotypique.

Soit *a* un allèle qui augmente la valeur du phénotype étudié ou qui augmente le risque d'être atteint d'une maladie. Un gène agit de façon additive sur un phénotype lorsque la valeur du phénotype d'un individu est proportionnelle au nombre d'allèles *a* que l'individu possède. Un effet dominant a lieu lorsque l'augmentation du phénotype est la même que l'individu possède une ou deux copies de l'allèle *a*. Un effet récessif se produit lorsque l'individu doit avoir deux allèles *a* afin d'avoir une valeur phénotypique plus élevée. Finalement, un effet génotypique nécessite un génotype particulier pour avoir un phénotype élevé: par rapport aux modèles précédents, dans le cas d'un SNP, un effet génotypique correspond à la situation dans laquelle les hétérozygotes ont une valeur phénotypique différente des homozygotes.

Tableau 3.1: Codifications génétiques pour le test de l'allèle a

Modèle \ Génotype	*aa*	*ab*	*bb*
Additif	2	1	0
Dominant	1	1	0
Récessif	1	0	0
Génotypique	0	1	0

Supposons que nous testons l'allèle *a* d'un SNP donné et que nous voulons codifier le génotype du sujet *j* de la famille *i*. La codification X_{ij} pour les quatre modèles génétiques est présentée au tableau 3.1. Par exemple, la codification additive s'obtient en assignant à X_{ij} le nombre d'allèles *a* que le sujet possède.

Le choix du modèle génétique est important: la puissance du test est plus grande si le modèle choisi correspond à la réalité. Malheureusement, le modèle réel est le plus souvent inconnu. Deux approches ont été proposées. D'une part, Laird (2006) note que le modèle additif se révèle adéquat même lorsque le gène n'agit pas sur le phénotype de manière additive. D'autre part, certains auteurs ont montré qu'il était plus puissant de tester tous les modèles génétiques plutôt que de tester un seul modèle, malgré la correction pour tests multiples plus sévère qui en découle (Freidlin et al., 2002). Nous avons examiné les deux approches dans les expériences que nous avons menées (chapitre 7).

3.4 Exemple

Nous présentons dans cette section un exemple artificiel afin d'illustrer le test FBAT. Examinons la famille nucléaire présentée à la figure 3.1. Étudions l'allèle 1 et considérons un modèle génétique additif. Les valeurs génotypiques X_{ij} pour les quatre enfants sont 2, 0, 1 et 1, respectivement. En supposant que la prévalence de la maladie dans la population étudiée est de 50%, un bon choix pour coder les phénotypes est d'assigner 0,5 aux individus malades et -0,5 à ceux qui ne le sont pas. Ainsi, les valeurs phénotypiques T_{ij} sont 0,5, -0,5, 0,5 et -0,5, respectivement. Nous avons toutes les

données pour calculer S:

$$\begin{aligned} S &= T_{11}X_{11}+T_{12}X_{12}+T_{13}X_{13}+T_{14}X_{14} \\ &= 0{,}5\cdot 2-0{,}5\cdot 0+0{,}5\cdot 1-0{,}5\cdot 1 \\ &= 1 \end{aligned}$$

L'espérance de S se calcule facilement puisque nous connaissons les génotypes des deux parents: chaque enfant devrait recevoir en moyenne un allèle 1.

$$E[S]=0{,}5\cdot 1-0{,}5\cdot 1+0{,}5\cdot 1-0{,}5\cdot 1=0$$

La statistique de test U est donc $U=S-E[S]=1$. Une valeur de U supérieure à 0 suggère que l'allèle 1 du SNP est associé à un risque accru de développer la maladie.

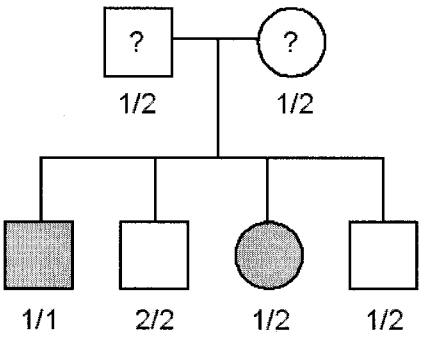

Figure 3.1: Exemple d'application du test FBAT

La famille nucléaire comprend deux parents et quatre enfants. Les carrés représentent des hommes, les cercles des femmes. Le phénotype est représenté à l'intérieur de la forme géométrique: le gris représente la maladie, le blanc l'absence de maladie et un point d'interrogation un statut inconnu. Les nombres sous les individus représentent leur génotype à un SNP. On remarque que le premier enfant en partant de la gauche est malade et a reçu l'allèle 1 de ses deux parents; le deuxième enfant n'est pas malade et a reçu l'allèle 2 de ses deux parents; le troisième enfant est malade alors que le quatrième ne l'est pas; ils ont tous deux reçu un allèle 1 d'un parent et un allèle 2 de l'autre parent (il est impossible de déterminer de quel parent provient chaque allèle).

3.5 Hypothèses nulles

FBAT permet de tester deux hypothèses nulles: « pas d'association ni de liaison » (notée

H_{01}) et « pas d'association en présence de liaison » (notée H_{02}; Laird, 2006). L'hypothèse alternative dans les deux cas est « présence d'association et de liaison ». La première hypothèse nulle est testée lorsque nous ne savons pas si le marqueur est lié au phénotype; à l'inverse, si nous savons que le marqueur est lié au phénotype, la deuxième hypothèse nulle est plus appropriée. Le test de H_{01} se fait sur la statistique de test Z calculée à partir des équations 3.1 à 3.3 décrites précédemment.

Par contre, le test de H_{02} requiert de tenir compte de la liaison entre les génotypes des enfants d'une même famille.

En effet, lorsque le marqueur est lié au phénotype, les transmissions génétiques d'un parent à ses enfants ne sont pas indépendantes. Pour illustration, supposons qu'un marqueur possède deux allèles (m_1 et m_2) et que le polymorphisme déterminant la maladie possède également deux allèles (p_1 et p_2). Supposons que les deux haplotypes d'un certain parent sont m_1p_1 et m_2p_2: l'allèle m_1 de ce parent se situe sur le même chromosome que l'allèle p_1 alors que l'allèle m_2 se situe sur le même chromosome que l'allèle p_2. Puisque le marqueur est lié au phénotype, le fait que le parent transmette son allèle m_1 à un enfant augmente à plus de ½ la probabilité de lui de transmettre son allèle p_1; à la limite, lorsque la fraction de recombinaison θ entre le marqueur et le polymorphisme tend vers 0, le fait de transmettre l'allèle m_1 implique la transmission de l'allèle p_1. Supposons que le polymorphisme p_1 augmente le risque de maladie de manière additive. Pour les enfants du parent étudié, l'allèle m_1 apparaîtra comme étant associé à la maladie, malgré le fait que le marqueur et le polymorphisme ne sont pas nécessairement en déséquilibre de liaison: la liaison donne l'impression d'association.

La solution proposée par Horvath et al. (2004) est d'ajuster la variance de S à l'équation 3.3.

Chapitre 4: Méthodes de correction actuelles

Nous rappelons d'abord quelques notions statistiques de base dans le cas d'un test statistique unique puis nous examinons les définitions de taux d'erreur lorsque plusieurs tests sont effectués. Nous sommes alors en mesure de définir formellement le problème que nous désirons résoudre. Nous décrivons ensuite les principales méthodes de correction pour tests multiples présentement utilisées. Finalement, nous récapitulons les caractéristiques de ces méthodes et indiquons en quoi elles ne conviennent pas au problème étudié.

4.1 Rappel statistique

Nous nous situons dans le contexte de procédures classiques de test d'hypothèses (*classical hypothesis test procedures,* Ostle et al., 1996, p. 230). Pour un problème donné, une hypothèse nulle, notée H_0, et une hypothèse alternative, notée H_a, sont posées. Un test statistique est choisi. Un seuil de significativité, noté α, est choisi. L'application du test statistique à un ensemble de données produit une statistique de test. À cette statistique de test correspond une valeur p, qui peut être calculée de façon analytique lorsque la distribution de la statistique de test sous l'hypothèse nulle est connue, ou de façon expérimentale. Lorsque la valeur p est inférieure ou égale au seuil α, l'hypothèse nulle H_0 est rejetée et l'hypothèse alternative H_a est acceptée. Lorsque la valeur p est supérieure au seuil α, l'hypothèse nulle n'est pas rejetée (Ostle et al., 1996, pp. 4, 5 et 235).

Dans le cas de tests d'association, la forme générale de l'hypothèse nulle est « le marqueur n'est pas associé au phénotype étudié » alors que l'hypothèse alternative est « le marqueur est associé au phénotype étudié ». Le test statistique que nous employons est le test FBAT (*family-based association test*). Tel que noté à la section 3.1, la statistique de test FBAT est approximativement distribuée selon une loi normale de moyenne 0 et de variance 1. La valeur p correspondant à cette statistique de test peut

être calculée de façon analytique en la comparant à la distribution normale.

La valeur p représente la probabilité d'observer une statistique de test aussi extrême ou plus extrême lorsque l'hypothèse nulle est vraie (Ostle et al., 1996, p. 233; Gigerenzer, 2004). En formule: $p = Pr(D \mid H_0)$, dans lequel D représente la statistique de test obtenue.

L'hypothèse nulle peut être vraie ou fausse et elle peut être rejetée ou ne pas être rejetée (tableau 4.1). Une erreur de type I est commise lorsqu'une hypothèse nulle vraie est rejetée: on parle de faux positif. Une erreur de type II est commise lorsqu'une hypothèse nulle fausse n'est pas rejetée: on parle de faux négatif. Un test statistique est dit fiable (*level-robust* selon l'expression de Westfall & Young, 1993) si les valeurs p rapportées par ce test proviennent d'une distribution uniforme entre 0 et 1 lorsque les hypothèses nulles testées sont vraies. Lorsque la distribution est biaisée vers de petites valeurs p, le test est libéral, alors qu'une distribution biaisée vers de grandes valeurs p indique un test conservateur. Lorsque le test est fiable, la probabilité qu'une erreur de type I soit commise correspond à α. La probabilité d'une erreur de type II est notée β.

Tableau 4.1: Hypothèse nulle et décision

H_0 / Action	***Rejetée***	***Pas rejetée***
Vraie	Faux positif	Vrai négatif
Fausse	Vrai positif	Faux négatif

L'expression $1 - \beta$ correspond à la puissance du test, c'est-à-dire à la probabilité de rejeter une hypothèse nulle fausse. Idéalement, nous aimerions que les taux d'erreur α et β s'approchent de 0: aucune erreur ne serait alors commise. Il y a toutefois un compromis à faire entre α et β: pour un ensemble de données et un test statistique choisi, diminuer une probabilité d'erreur augmente l'autre probabilité d'erreur. Si nous souhaitons diminuer un taux d'erreur sans augmenter l'autre, des avenues envisageables consistent à augmenter la taille de notre échantillon (Ostle et al., 1996), utiliser un test statistique plus puissant ou utiliser un sous-ensemble de données plus homogène

(Hauser et al., 2004; Hamet et al., 2005; Pausova et al., 2005). Un test conservateur est un test de faible puissance, à cause du compromis entre α et β.

4.2 Taux d'erreur au sein de la famille de tests

Lorsque plusieurs hypothèses nulles sont testées, par exemple lorsqu'on teste l'association entre plusieurs marqueurs génétiques et un phénotype, des mesures d'erreur autres que l'erreur de type I doivent être utilisées et contrôlées. En effet, lorsque n hypothèses indépendantes sont testées, la probabilité de faire au moins une erreur de type I est $1-(1-\alpha)^n$. Par exemple, pour $\alpha = 0,05$ et $n = 10$, la probabilité de rejeter au moins une hypothèse nulle vraie est 0,40. Un seuil de 0,05 appliqué à chaque test n'est donc pas approprié dans le cas de tests multiples. Le problème est particulièrement aigu dans les études d'association génétiques: lorsque l'association entre 50 000 marqueurs et un phénotype est testée, on peut s'attendre à ce qu'environ 5% des tests, soit 2500, rapportent une valeur p inférieure à 0,05, même si aucun des marqueurs n'est associé au phénotype. Une étude qui rapporterait 2500 associations « significatives » contiendrait probablement beaucoup plus de faux positifs que de vrais positifs et serait de peu d'utilité pour le monde scientifique.

Plusieurs généralisations de l'erreur de type I applicables aux tests multiples ont été développées (Storey, 2002). Le taux d'erreur que nous souhaitons contrôler dans la présente étude est le taux d'erreur au sein de la famille de tests (*family-wise error rate*, FWER). Ce taux est défini comme étant la probabilité de rejeter au moins une hypothèse nulle vraie. Contrôler le FWER à un taux α consiste à s'assurer que la probabilité de rejeter au moins une hypothèse nulle vraie est au plus α. En se servant du tableau 4.2, on définit: FWER = Pr(V>0).

Il existe deux variantes du FWER: le taux d'erreur au sein de la famille de tests lorsque toutes les hypothèses nulles sont vraies (noté FWEC) et le taux d'erreur au sein de la famille de tests lorsqu'un sous-ensemble des hypothèses nulles est vrai (noté FWEP; Westfall & Young, 1993). De façon plus formelle:

- FWEC = Pr(rejeter au moins une H_{0i} | toutes les H_{0i} sont vraies)
- FWEP = Pr(rejeter au moins une H_{0i}, i e $\{j_1, ..., j_t\}$| $H_{0j1}, ...,$ H_{0jt} sont vraies)

Le contrôle de FWEP est plus rigoureux que le contrôle de FWEC, car le taux d'erreur doit être contrôlé peu importe quel sous-ensemble des hypothèses nulles est vrai. Contrôler le FWEP se dit contrôler *fortement* le FWER alors que contrôler le FWEC se dit contrôler *faiblement* le FWER (Westfall & Young, 1993).

Tableau 4.2: Tests multiples (adapté de Benjamini & Hochberg, 1995)

H_0 / Action	***Rejetées***	***Non rejetées***	***Total***
Vraies	V	U	m_0
Fausses	S	T	$m - m_0$
Total	R	m - R	m

4.3 Autres taux d'erreur

Lorsqu'on contrôle le FWER, les valeurs p corrigées dépendent non seulement du test considéré mais également des autres tests effectués dans la même expérience. Ainsi, un marqueur d'intérêt peut être déclaré significatif ou non selon les marqueurs testés avec lui. Cela peut paraître non-intuitif. Plutôt que de contrôler le FWER, il est possible de contrôler la probabilité qu'une association rapportée soit fausse (*false positive report probability*, FPRP). Ce taux d'erreur ne dépend pas des marqueurs testés avec le msrqueur d'intérêt.

On peut aussi vouloir contrôler le taux de fausses découvertes (*false discovery rate*, FDR). Ce taux dépend du nombre de marqueurs testés simultanément, mais son contrôle peut être moins sévère que le contrôle du FWER, ce qui peut augmenter le nombre d'associations significatives rapportées. Nous examinons tour à tour la FPRP et le FDR.

4.3.1 Probabilité qu'une association rapportée soit fausse

La FPRP a été proposée par Wacholder et al. (2004). Ce taux d'erreur représente la probabilité qu'il n'y a en réalité aucune association étant donné un résultat statistiquement significatif. En formule: FPRP = $Pr(H_0 \mid p \leq \alpha)$. La FPRP dépend de trois éléments: la valeur p observée, la puissance du test et la probabilité à priori que l'association soit vraie. La puissance du test dépend à son tour de la taille de l'échantillon, de la fréquence du polymorphisme, de la taille de son effet (*effect size*), de son modèle génétique et du LD entre le marqueur testé et le polymorphisme qui cause la maladie.

De façon analogue, le Wellcome Trust Case Control Consortium (2007) calcule la chance (*odds*) que le résultat soit vrai:

$$Chance\,à\,posteriori = \frac{Chance\,à\,priori \times (1-\beta)}{\alpha} \tag{4.1}$$

Rappelons que $1-\beta$ et α représentent respectivement la puissance et le seuil de significativité du test. La chance qu'une association statistiquement significative soit vraie augmente lorsque la chance à priori augmente, la puissance du test augmente, et le seuil de significativité diminue. On constate donc que la valeur p rapportée par le test statistique détermine en partie la chance d'une vraie association mais que la puissance et la chance à priori ont également un rôle. Par contre, le nombre de tests effectués dans une expérience n'influence pas la chance à posteriori.

La méthode FPRP procède en quatre étapes:

1. Fixer un seuil de FPRP tel que les associations dont la FPRP sera plus petite ou égale à ce seuil seront déclarées « intéressantes ». Ceci est analogue à fixer un seuil α tel que les valeurs p inférieures ou égales sont déclarées significatives.
2. Déterminer la probabilité à priori de l'hypothèse alternative de chaque test. Ces probabilités peuvent être différentes les unes des autres, par exemple un

marqueur pour lequel plusieurs associations significatives ont déjà été publiées aurait une probabilité à priori plus élevée qu'un marqueur quelconque. Plusieurs probabilités à priori peuvent être fixées afin de vérifier la sensibilité de la méthode.

3. Calculer la puissance des tests en fonction de la fréquence de l'allèle mineur, de la taille de l'échantillon, de la taille de l'effet du polymorphisme, du modèle génétique supposé et du LD supposé entre le marqueur et le polymorphisme causant la maladie.
4. Effectuer les tests d'association, calculer la FPRP de chacun et déterminer s'ils sont « intéressants ».

La FPRP répond à la question « Quelle est la probabilité que l'hypothèse nulle soit vraie, étant donné qu'elle a été rejetée? » Selon certains auteurs, la réponse à cette question peut être d'une plus grande utilité que la connaissance seule de la valeur p.

Berger et Sellke (1987) montrent, pour une large gamme de probabilités à priori de l'hypothèse nulle, que des valeurs p relativement faibles peuvent correspondre à des probabilités à posteriori de l'hypothèse nulle relativement élevées. Par exemple, en faisant certaines suppositions, dont une probabilité à priori de 0,5, une valeur p de 0,05 amène à une probabilité à posteriori de 0,23. Ainsi, malgré le fait que la valeur p soit statistiquement significative pour un seuil nominal de 0,05, il y a tout de même près d'une chance sur quatre pour que l'hypothèse nulle soit vraie.

Ioannidis (2005) examine les probabilités à posteriori qu'une association soit vraie dans différents types d'études. Par exemple, des études pour lesquelles la probabilité à priori d'association est faible (0,1) et pour lesquelles la puissance est forte (0,80) augmentent la probabilité à posteriori à seulement 0,2 lorsque la valeur p obtenue est de 0,05. Lorsqu'on diminue la puissance à 0,20, la probabilité à posteriori n'est alors que de 0,12, à peine mieux qu'avant le début de l'étude.

Selon Gigerenzer (2004), la grande majorité des étudiants, chercheurs et professeurs ont

tendance à interpréter la valeur p comme étant la probabilité de l'hypothèse nulle. Cette erreur d'interprétation est partagée par plusieurs auteurs de manuels statistiques.

Nous en concluons que la valeur p seule n'est pas nécessairement un bon indicateur de la véracité de l'hypothèse nulle. Aussi, il est fréquent que la valeur p soit mal interprétée. Dans ces conditions, l'utilisation d'une mesure comme la FPRP fournit une meilleure appréciation de la véracité de l'hypothèse nulle et évite les erreurs d'interprétation.

Il y a toutefois un prix à payer pour connaître la FPRP: il faut spécifier la probabilité à priori de chaque hypothèse et calculer la puissance du test statistique. La gamme de probabilités à priori raisonnables peut être grande (Westfall & Young, 1993, p. 22). Aussi, la puissance est calculée sous de nombreuses suppositions. Il en résulte que deux chercheurs peuvent arriver à des conclusions fort différentes.

Malgré tout, une association réelle dont l'effet est important calculée sur un échantillon de grande taille pourra donner une faible FPRP pour une large gamme de probabilités à priori et de puissances supposées.

4.3.2 Taux de fausses découvertes

Le taux de fausses découvertes est défini comme étant la proportion d'hypothèses nulles vraies rejetées parmi toutes les hypothèses nulles rejetées (Benjamini & Hochberg, 1995). Pour tenir compte de la situation dans laquelle aucune hypothèse nulle n'est rejetée, le taux positif de fausses découvertes (*positive false discovery rate*, pFDR) est défini comme étant la proportion d'hypothèses nulles vraies rejetées parmi toutes les hypothèses nulles rejetées, lorsqu'au moins une hypothèse nulle a été rejetée (Dudbridge et al. 2006; Storey, 2002). En se servant du tableau 4.2, on définit: FDR = E(V/R).

Le contrôle du taux de fausses découvertes convient aux études exploratoires dans lesquelles un grand nombre de tests sont effectués afin de déterminer un ensemble plus restreint d'hypothèses à être investiguées de façon plus approfondie (Benjamini & Hochberg, 1995; Storey, 2002; Efron, 2004).

Le taux de fausses découvertes est attrayant car en général on obtient une puissance plus grande en contrôlant ce taux plutôt qu'en contrôlant le FWER (Storey, 2002).

Le FDR souffre par contre de deux difficultés. Premièrement, il n'est pas clair comment FDR et pFDR tiennent compte des corrélations dans les expériences que nous avons effectuées. Entre autres, les définitions de Benjamini et Hochberg (1995) et Storey (2002) se limitent au cas où tous les tests sont indépendants. Benjamini et Yekutieli (2001) se sont attardés aux cas où les tests étaient corrélés. Ils ont prouvé que le FDR était applicable lorsque les tests satisfont une propriété de dépendance positive (*positive regression dependency on each one from a subset* I_0, abrévié PRDS). Ils donnent quelques exemples de tests qui satisfont cette propriété mais reconnaissent que la question n'est pas résolue pour d'autres tests, entre autres pour certains tests bilatéraux. Sabatti et al. (2003) ont appliqué la méthode aux tests d'association dans lesquels les marqueurs peuvent être en LD. Ils affirment que la propriété PRDS « apparaît » être satisfaite, d'après une justification intuitive et quelques simulations, mais ne le démontrent pas formellement. Notons que nos expériences contiennent plusieurs formes de dépendances: entre marqueurs, entre phénotypes, entre modèles génétiques et entre sous-groupes d'individus. La validité du FDR appliqué à nos expériences est donc incertaine pour le moment.

Deuxièmement, le contrôle de ce taux d'erreur n'offre pas de garantie: ce n'est pas une probabilité qui est contrôlée, mais une proportion attendue. En particulier la variation de la proportion réelle de faux positifs est très grande lorsque la puissance des tests est faible (Storey, 2002). Alors que le contrôle du FWER à un niveau α nous garantit que la probabilité qu'au moins une erreur de type I ait été commise est au maximum α, le contrôle du pFDR au niveau α nous indique seulement que la proportion *attendue* de faux positifs est α. La proportion réelle peut être très différente d'une expérience à l'autre. De plus, les formulations originales de FDR et de pFDR ne permettent pas de quantifier ni de contrôler cette variabilité. Ainsi, un chercheur ne sait pas si dans l'expérience qu'il mène sa proportion de fausses découvertes est effectivement α ou bien si elle est beaucoup plus grande ou beaucoup plus petite.

4.4 Définition du problème

Nous avons besoin d'une méthode de correction pour tests multiples qui s'applique aux études d'association génomique familiale. La méthode doit:

1. contrôler fortement le taux d'erreur au sein de la famille de tests (FWER);
2. supporter les familles de grandes tailles;
3. supporter les patrons de génotypes manquants arbitraires, en particulier les parents qui ne sont pas génotypés;
4. supporter les distributions phénotypiques arbitraires, entre autres les phénotypes qui ne montrent aucune variabilité (c'est-à-dire que le phénotype étudié a la même valeur chez tous les individus phénotypés);
5. s'appliquer à tout test d'association génomique familiale;
6. supporter un grand nombre de SNPs (des dizaines de milliers) et de tests statistiques (des centaines de milliers);
7. être puissante.

Nous sommes maintenant en mesure d'examiner les méthodes de correction pour tests multiples présentement utilisées et vérifier si elles respectent les spécifications.

4.5 Correction de Bonferroni

La correction proposée par Carlo Emilio Bonferroni est certainement la méthode de correction pour tests multiples la plus simple à implanter. Lorsque n hypothèses sont testées, la valeur p corrigée correspondant à une valeur p observée est donnée par:

$$p_{corrigée} = min(1,\ n \cdot p_{observée}) \tag{4.2}$$

Cette valeur p corrigée est comparée au seuil de significativité α. Le minimum est utilisé afin de borner la valeur p corrigée entre 0 et 1. La correction de Bonferroni contrôle fortement la FWER. De plus, lorsque les n tests statistiques sont indépendants et sont

suffisamment puissants, cette correction est exacte: aucune méthode qui contrôle fortement le FWER ne peut produire de meilleures valeurs p corrigées.

Par contre, lorsque les tests sont corrélés, la correction de Bonferroni peut être très conservatrice. Dans les tests d'association entre marqueurs et phénotypes, au moins quatre types de corrélations existent:

1. entre marqueurs (à cause du déséquilibre de liaison);
2. entre phénotypes;
3. entre modèles génétiques (Freidlin et al., 2002);
4. entre sous-groupes étudiés.

Pour illustrer la corrélation entre SNPs, nous avons calculé, dans les données étudiées (voir la section 7.1.2), le nombre de SNPs qui sont en LD avec au moins un autre SNP. En tout, 54 524 SNPs autosomiques polymorphiques sont à notre disposition. Les résultats sont présentés au tableau 4.3 pour différents seuils de LD.

Tableau 4.3: LD entre SNPs

Seuil minimal de LD (r^2)	***Nombre de SNPs en LD avec au moins un autre SNP***	***% de SNPs***
1,0	6 616	12,1
0,8	19 847	36,4
0,6	24 684	45,3
0,4	30 162	55,3
0,2	37 879	69,5
0,0	54 524	100,0

Pour un seuil de r^2 plus grand ou égal à 0,4, plus de la moitié des SNPs sont en LD avec au moins un autre SNP. Pour un seuil plus rigoureux de 0,8, la proportion est du tiers des SNPs. Le déséquilibre de liaison est bien présent dans nos données, ce qui nous amène à prévoir que la correction de Bonferroni sera conservatrice.

4.6 Correction par simulations dans les études cas-témoins

Les valeurs p calculées dans les études cas-témoins peuvent être corrigées par une méthode de rééchantillonnage fort simple: il suffit de permuter les phénotypes sans modifier les génotypes (Dudbridge et al., 2006). Ainsi, la corrélation entre les SNPs est conservée. Ces simulations sont relativement rapides, de l'ordre de quelques secondes par réplicat. Les méthodes de rééchantillonnage sont décrites en détail à la section 5.1.

Nous aimerions appliquer ce type de correction aux études familiales. Malheureusement, cela pose plusieurs problèmes.

Premièrement, une permutation des valeurs phénotypiques suppose qu'il existe de la variabilité dans le phénotype observé. Or, dans certaines études de trios, qui représentent un cas particulier des études familiales, seuls les enfants affectés par la maladie sont génotypés, alors que le phénotype des parents n'est pas mesuré. Il n'y a alors aucune variabilité, donc aucune possibilité de permuter les phénotypes.

Deuxièmement, en permutant librement les phénotypes entre les individus, la corrélation phénotypique entre individus généalogiquement rapprochés est perdue. La variabilité du phénotype est due en partie aux gènes et en partie à l'environnement. L'héritabilité est définie comme étant le rapport de la variabilité due aux gènes à la variabilité totale du phénotype. L'héritabilité est communément calculée avant d'entreprendre des études génétiques: si un phénotype ne montre aucune héritabilité, il est inutile d'essayer de trouver des causes génétiques. La perte d'héritabilité dans les phénotypes permutés ne pose pas de problèmes aux études cas-témoins car elles contiennent des sujets qui sont peu reliés entre eux, par conception de l'étude. Par contre, dans un contexte d'études familiales, la perte d'héritabilité peut rendre le test conservateur dans certains cas et libéral dans d'autres cas, notamment pour les raisons suivantes:

- Un test d'association qui suppose que les sujets et leurs génotypes et phénotypes sont indépendants les uns des autres (un test T, par exemple) pourrait être extrêmement libéral (valeurs p corrigées plusieurs ordres de grandeur trop basses; Newman et al., 2001). Une permutation libre des phénotypes ne

corrigerait pas cette libéralité. Bien que le but d'une méthode de correction pour tests multiples n'est pas de pallier les lacunes du test d'association sous-jacent, y arriver rend la méthode plus robuste.

- L'hypothèse nulle « pas d'association en présence de liaison » de FBAT est testée en corrigeant la variance de S pour les liens au sein d'une fratrie. Cette correction s'avère trop sévère lorsque les phénotypes sont permutés librement.

La permutation libre des phénotypes n'est donc pas robuste dans les études familiales: des données qui contiennent peu de variabilité phénotypique, un test d'association sous-jacent qui ne tient pas compte des relations généalogiques et le test de l'hypothèse nulle « pas d'association en présence de liaison » par FBAT sont trois situations pour lesquelles la permutation libre des phénotypes s'avère conservatrice ou libérale.

4.7 Correction dans les études de trios

Un type d'étude très répandu est celui des trios, dans lequel les deux parents et un enfant sont enrôlés. Les trois membres du trio sont génotypés et seul l'enfant est phénotypé. Un test TDT, comme le test FBAT, est ensuite appliqué.

Plusieurs auteurs ont proposé des méthodes de rééchantillonnage pour corriger les valeurs p provenant de ces tests. La méthode de Lazzeroni et Lange (1998) traite un bloc d'haplotypes à la fois, mais la méthode peut être appliquée aisément aux marqueurs individuellement. Pour chaque parent, un de ses deux haplotypes est transmis à l'enfant avec une probabilité ½. Le test TDT est alors calculé pour chaque bloc d'haplotypes à partir des génotypes simulés et des phénotypes observés. Des approches similaires ont été implantées dans les logiciels MENDEL (Lange et al., 2005), PLINK (Purcell et al., 2007) et Haploview (Barrett, 2005).

Cette approche requiert le génotypage des deux parents. En effet, il faut connaître les génotypes parentaux pour savoir quel allèle n'a pas été transmis à l'enfant. Cela rend l'approche inutilisable dans les études qui contiennent de nombreux parents non génotypés.

4.8 Simulation des études familiales

Le logiciel *Merlin* (Abecasis, 2002) a été conçu pour effectuer des analyses de liaison sur des données familiales. *Merlin* peut générer des génotypes qui respectent les fréquences de recombinaison et les fréquences alléliques originales, les patrons de génotypes manquants et de façon grossière le déséquilibre de liaison.

Merlin pourrait être utilisé pour générer des haplotypes qui seraient ensuite utilisés par *FBAT*. Par contre, la complexité algorithmique de *Merlin* (en terme de temps de calcul et de mémoire) est exponentielle par rapport au nombre de non-fondateurs dans une famille. (Un fondateur est défini comme étant un individu pour lequel les deux parents sont inconnus.) De façon pratique, sur un ordinateur possédant 1,5 Go de mémoire vive, *Merlin* peut traiter une famille de 10 à 12 membres en un temps raisonnable. Pour des familles plus grosses, il est possible de répartir ses membres en plusieurs sous-familles afin de respecter les contraintes de temps de calcul et de mémoire. Par contre, les haplotypes ainsi simulés risquent de ne pas correspondre tout à fait à la réalité et il est difficile d'estimer l'impact de cette répartition.

Nous ne connaissons pas d'autres méthodes de simulation d'études familiales qui s'appliquent aux tests d'association. Les méthodes actuelles ne s'appliquent qu'aux tests de liaison: *SOLAR* (Almasy & Blangero, 1998) simule un seul marqueur parfaitement informatif mais dont les fréquences alléliques ne reflètent pas les fréquences alléliques originales, alors que *SIBPAL* (S.A.G.E., 2006) permute la proportion de partage d'allèles *identical by descent* pour un seul marqueur. Dans les deux cas, un seul marqueur est simulé à la fois: le LD entre marqueurs n'est pas simulé.

4.9 Sommaire des méthodes existantes

En résumé, les méthodes existantes de correction pour tests génétiques multiples ne sont pas prometteuses: elles sont soit peu puissantes, soit inapplicables à notre problème.

- La correction de Bonferroni risque d'être peu puissante car elle ne tient pas

compte des corrélations entre les tests.

- La simulation d'études cas-témoins ne tient pas compte de l'héritabilité du phénotype et n'est donc pas applicable aux études familiales. De plus, elle suppose une variabilité dans les valeurs phénotypiques qui n'est pas nécessairement présente dans les études familiales.
- La simulation d'études de trios requiert le génotypage des parents.
- La simulation des études familiales est limitée: elle ne supporte pas les familles de grande taille et elle ne modélise que grossièrement le déséquilibre de liaison. L'impact de ces limites est difficile à évaluer.

Notons finalement que certains chercheurs ont tenté de déterminer un nombre effectif de tests qui tiendrait compte de la corrélation entre les tests. Il suffirait ensuite d'appliquer la correction de Bonferroni en multipliant les valeurs p observées par ce nombre effectif plutôt que par le nombre total de tests. Nyholt (2004) a ainsi proposé de calculer le nombre effectif de tests en fonction de la variance des valeurs propres de la matrice de corrélation entre SNPs. Par contre, Salyakina et al. (2005) ont démontré théoriquement et expérimentalement que cette correction n'est pas fiable: elle est libérale lorsque le LD augmente et conservatrice lorsque les SNPs sont regroupés en blocs d'haplotypes.

Il est donc nécessaire de développer une méthode qui contrôle fortement le FWER de tests multiples d'association génomique familiale et qui offre une bonne puissance tout en supportant de grandes familles, des parents non génotypés et un LD fin.

Chapitre 5: Méthode proposée

La méthode que nous proposons vise à respecter autant que possible les caractéristiques des données réelles afin qu'elle soit fiable et puissante. Elle est basée sur le rééchantillonnage (*resampling*; Westfall & Young, 1993).

Dans ce chapitre nous présentons d'abord les méthodes de rééchantillonnage. Nous décrivons ensuite la méthode proposée. Nous examinons ses caractéristiques afin de nous assurer qu'elle résoud le problème posé au chapitre précédent.

5.1 Méthodes de rééchantillonnage

Les méthodes de rééchantillonnage calculent une valeur p en n'utilisant que les données disponibles, sans les comparer à une distribution théorique. Elles y arrivent en construisant plusieurs échantillons à partir des données originales. Ces échantillons peuvent être créés avec remise (*bootstrap*) ou sans remise (permutation). La statistique de test est d'abord calculée sur les données originales (i.e. sans modification). La statistique de test est ensuite calculée sur chaque échantillon. La valeur p empirique est la proportion d'échantillons qui ont produit une statistique de test aussi extrême ou plus extrême que la statistique de test originale.

Les méthodes de rééchantillonnage sont utilisées dans deux contextes: pour tester une seule hypothèse ou pour en tester plusieurs. Lorsqu'une seule hypothèse est testée, le rééchantillonnage permet de calculer une valeur p sans faire de suppositions sur la distribution des données, ce qui donne des tests moins libéraux ou plus puissants que les tests paramétriques lorsque leurs suppositions ne sont pas respectées (Westfall & Young, 1993, p. 13; Seda et al., 2008). Lorsque plusieurs hypothèses sont testées, le rééchantillonnage permet de calculer une valeur p qui tient compte de tous les tests qui ont été effectués en incorporant la corrélation entre les tests, ce qui le rend en général plus puissant que les méthodes de correction qui ne considèrent que le nombre de tests

effectués.

Afin d'être fiables, les méthodes de rééchantillonnage doivent posséder deux propriétés (Hall & Wilson, 1991):

1. le rééchantillonnage doit se faire sous l'hypothèse nulle;
2. la statistique de test utilisée doit être pivotale.

La première condition est directement liée à la définition d'une valeur p, soit la probabilité d'observer un résultat aussi extrême ou plus extrême *lorsque l'hypothèse nulle est vraie*. Les données doivent donc être simulées comme si l'hypothèse nulle était vraie. Concernant la deuxième condition, une statistique de test est pivotale lorsqu'elle ne dépend pas des paramètres de la distribution dont proviennent les données étudiées lorsque l'hypothèse nulle est vraie. Par exemple, la statistique de test bien connue $T=(\bar{Y}-\mu)/(s/\sqrt{n})$ est pivotale car elle est distribuée selon une loi normale de moyenne 0 et de variance 1, peu importe la distribution ayant généré Y (grâce au théorème de la limite centrale); en particulier, T ne dépend ni de la moyenne ni de la variance de la distribution ayant généré Y.

Dans le contexte d'hypothèses multiples, une troisième condition doit être satisfaite afin de contrôler fortement le FWER (Westfall & Young, 1993, p. 43; Ge et al., 2003, p. 12):

3. les valeurs p non corrigées doivent posséder la propriété de pivot du sous-ensemble (*subset pivotality*).

Cette troisième condition permet d'effectuer les simulations en supposant toutes les hypothèses nulles vraies. Lorsque cette condition n'est pas respectée, la méthode ne contrôle que faiblement le FWER. La propriété de pivot du sous-ensemble est définie comme suit. Soit m le nombre de tests. Une distribution de valeurs p non corrigées (P_1, P_2, ..., P_m) possède la propriété de pivot du sous-ensemble si pour tous les sous-ensembles K de {1, 2, ..., m}, la distribution conjointe de $\{P_i \mid i \in K\}$ est la même peu importe que toutes les hypothèses nulles soient vraies ou seulement les K. En d'autres termes, les valeurs p du sous-ensemble considéré ne dépendent pas de la véracité

d'hypothèses qui ne font pas partie du sous-ensemble considéré.

Trois sources d'erreurs peuvent fausser les résultats produits par une méthode de rééchantillonnage:

1. l'erreur de simulation;
2. l'erreur due à l'utilisation d'un générateur de nombres pseudo-aléatoires;
3. l'erreur due à l'utilisation des données collectées pour générer les échantillons.

L'erreur de simulation représente le fait que la valeur p est estimée à partir d'un nombre fini d'échantillons et n'est donc pas tout à fait égale à la valeur p réelle. Cette erreur diminue au fur et à mesure que le nombre de simulations augmente. Nous quantifions cette erreur à la prochaine sous-section (Calcul de l'erreur standard).

Idéalement, nous devrions utiliser un générateur de nombres parfaitement aléatoires. En pratique, nous utilisons un générateur de nombres *pseudo*-aléatoires, ce qui cause une erreur. L'erreur peut être réduite en utilisant un générateur qui possède de bonnes propriétés. Nous examinons les propriétés d'un bon générateur à la section 6.5.1 et décrivons les générateurs utilisés dans notre implantation.

Nous générons les échantillons à partir des données collectées plutôt que des données de toute la population. Les échantillons générés ressemblent donc aux données collectées plutôt qu'à la population, ce qui pourrait limiter l'inférence à la population. Par contre, l'impact de cette source d'erreur diminue lorsque la taille de l'ensemble de données augmente. Au chapitre 7 nous décrivons la cohorte étudiée.

En résumé, l'impact des trois sources d'erreurs peut être diminué en effectuant un nombre élevé de simulations, en utilisant un bon générateur de nombres pseudo-aléatoires et en utilisant un jeu de données de taille adéquate.

5.1.1 Calcul de l'erreur standard

Les valeurs p rapportées par une méthode de simulation approximent la valeur p réelle.

Plus le nombre de simulations augmente, plus l'approximation devient précise. Les résultats des simulations suivent une distribution binomiale: le résultat d'une simulation peut être plus extrême ou égal au résultat original (succès) ou moins extrême (insuccès). Soit q une valeur p empirique calculée à partir de N simulations. Son erreur standard peut être calculée de façon asymptotique lorsque le nombre de succès est supérieur ou égal à 5 et que le nombre d'insuccès est également supérieur ou égal à 5, grâce au théorème de la limite centrale (Ostle et al., 1996; Westfall & Young, 1993):

$$se=\sqrt{(\frac{q\cdot(1-q)}{N})} \tag{5.1}$$

Un intervalle de confiance à 95% peut être construit (Westfall & Young, 1993, p. 41):

$$ci=[\,q-1{,}96\cdot se\,;q+1{,}96\cdot se\,] \tag{5.2}$$

Lorsque l'approximation normale n'est pas valide, l'intervalle de confiance doit être calculé par des méthodes numériques (Ott, 1991, pp. 49-50). La librairie *binom* de l'environnement statistique *R* offre plusieurs façons de calculer l'intervalle de confiance.

Rappelons qu'un intervalle de confiance à x% signifie qu'à long terme, x% des intervalles de confiance contiendront la valeur réelle (Ostle et al., 1996, p. 188).

L'intervalle de confiance peut être utile pour savoir si suffisamment de simulations ont été effectuées. Ainsi, si après un certain nombre de simulations l'intervalle de confiance comprend le seuil de significativité α, les simulations devraient se poursuivre. Par contre, si l'intervalle de confiance ne comprend pas α, les simulations peuvent cesser (Westfall & Young, 1993, p. 117). Selon Ott (1991, p. 191) un résultat ne devrait être déclaré significatif que lorsque la borne supérieure de l'intervalle de confiance d'une valeur p empirique est inférieure à α.

5.2 Description de la méthode

Nous aimerions développer une méthode qui se rapproche de la méthode de simulation

des études cas-contrôles, décrite à la section 4.6. Nous ne pouvons cependant pas simplement permuter les phénotypes: l'héritabilité serait perdue et la méthode ne serait pas applicable aux études conventionnelles de trios. Nous ne pouvons pas non plus permuter les haplotypes entre les individus: une des lois de Mendel stipule qu'un individu reçoit, à chaque locus, un allèle de chacun de ses parents. Nous pouvons par contre simuler les haplotypes de telle façon que leurs *propriétés* originales[2] soient conservées *et* que tous les génotypes respectent les lois de Mendel. Les principales propriétés que nous voulons respecter sont le déséquilibre de liaison entre les SNPs , les fréquences alléliques et les patrons de génotypes manquants.

Conceptuellement, la méthode que nous proposons procède selon les étapes décrites à la figure 5.1.

1. Calculer l'association entre les génotypes et phénotypes observés
2. Phaser les génotypes observés
3. Calculer le déséquilibre de liaison entre toutes les paires de SNPs
4. RÉPÉTER *N* fois:
5. Simuler deux haplotypes par fondateur
6. Transmettre les haplotypes des fondateurs à leurs descendants
7. Calculer l'association entre les génotypes simulés et les phénotypes originaux
8. Calculer les valeurs p corrigées
9. Déclarer statistiquement significatives les associations pour lesquelles la valeur p corrigée est plus petite ou égale au seuil α pré-établi

Figure 5.1: Méthode de correction pour tests multiples proposée

À l'étape 2, il est nécessaire de phaser les génotypes observés afin de calculer le LD entre les SNPs (étape 3). Aux étapes 5 et 6, les génotypes sont générés sans tenir compte des phénotypes. Ainsi, à chaque simulation, l'hypothèse nulle « il n'y a pas d'association entre les phénotypes et les marqueurs » est vraie.

Afin que notre simulation soit valide, nous testons la même hypothèse nulle aux étapes 1 et 7 et nous utilisons les mêmes paramètres, en particulier le modèle génétique et le

2 Nous utilisons les adjectifs « observé » et « original » pour décrire les données réelles: soit les phénotypes et les génotypes mesurés chez les sujets de notre population d'étude, soit les résultats provenant de l'association entre les génotypes réels et les phénotypes réels.

nombre minimum de familles requis pour effectuer le test.

À chaque simulation, nous conservons la meilleure valeur p parmi tous les tests effectués.

À l'étape 8, la valeur p corrigée (ou « ajustée ») pour un test est obtenue en divisant le nombre de simulations ayant rapporté une meilleure statistique de test FBAT que celle observée à ce test par le nombre de simulations (*N*).

On peut considérer que la méthode en est une de *bootstrap*, car les génotypes à chaque locus sont générés avec remise à partir du bassin d'allèles de la cohorte étudiée.

À l'étape 6, les haplotypes simulés des fondateurs sont transmis à leurs descendants. La méthode de *gene-dropping* (MacCluer, 1986; Tremblay et al., 2003; Jung, 2006) est tout indiquée pour réaliser cette transmission. Dans sa formulation originale, le *gene-dropping* considère un locus à la fois. Elle travaille sur une structure familiale connue mais n'utilise pas les données phénotypiques. Elle simule la transmission des allèles d'un gène des parents aux enfants selon les lois mendéliennes. Pour un locus autosomique, elle assigne deux allèles à chaque fondateur en tenant compte des fréquences alléliques spécifiées. Ensuite, les fondateurs transmettent un allèle à chacun de leurs enfants; chaque allèle d'un fondateur a une probabilité ½ d'être transmis à chacun de ses enfants. La transmission à un enfant est indépendante de la transmission à un autre enfant et la transmission provenant d'un parent est indépendante de la transmission provenant de l'autre parent. Lorsqu'un enfant a reçu ses deux allèles, il peut à son tour les transmettre à ses propres enfants.

Le *gene-dropping* suppose qu'il n'y a aucune force extérieure qui agit au locus en question: principalement, l'accouplement est aléatoire, il n'y a pas de pression sélective et il n'y a pas de mutation. En d'autres termes, le *gene-dropping* suppose que le locus est en équilibre de Hardy-Weinberg.

5.3 Contrôle fort de FWER

Nous avons affirmé à la section 4.2 que nous désirions contrôler fortement le taux d'erreur au sein de la famille de tests (FWER), c'est-à-dire la probabilité de rejeter au moins une hypothèse nulle vraie, peu importe le sous-ensemble d'hypothèses nulles vraies. Pour qu'une méthode de rééchantillonnage contrôle fortement le FWER, nous avons noté à la section 5.1 qu'elle doit respecter trois conditions: les simulations devaient se faire sous l'hypothèse nulle, le test de base doit être pivotal et les valeurs p originales doivent posséder la propriété de pivot du sous-ensemble.

La première condition (simuler sous l'hypothèse nulle) est respectée par construction: les génotypes sont assignés aux fondateurs puis transmis à leurs descendants *sans prendre en compte les phénotypes des individus*. Il n'y a donc aucune association dans les données simulées, comme le veut l'hypothèse nulle.

La statistique de test rapportée par *FBAT* ($Z = U/\sqrt{V}$) est approximativement distribuée selon une loi normale de moyenne 0 et de variance 1 (section 3.1). En d'autres termes, la moyenne et la variance de la distribution observée n'influencent pas la distribution de la statistique de test, lorsque le nombre d'observations est suffisamment grand (théorème de la limite centrale). La condition de pivot est donc satisfaite.

Dans le cas de tests d'association entre SNPs et phénotypes, la condition de pivot du sous-ensemble est également satisfaite: la valeur du test FBAT à un SNP ne dépend que de ce SNP et du phénotype. Ainsi, le fait que l'hypothèse nulle concernant un autre SNP soit vraie ou fausse ne change rien à la valeur p du test FBAT à ce SNP.

Notre méthode respecte les trois conditions stipulées par Westfall et Young (1993) et devrait donc être fiable. Nous le vérifions de façon expérimentale à la section 6.5.4.

5.4 Caractéristiques

Voici les principales caractéristiques de la méthode proposée:

- *Supporte les familles de grande taille*. Certains logiciels qui traitent les familles

sont de complexité exponentielle par rapport au nombre de non-fondateurs dans la famille. En pratique, ces logiciels débordent de la mémoire vive disponible ou nécessitent un temps de calcul déraisonnable lorsque la famille compte plus d'une dizaine d'individus. Or, certaines familles de notre population d'étude sont beaucoup plus grosses. En comparaison, tous les algorithmes utilisés par la méthode proposée sont de complexité linéaire par rapport au nombre d'individus par famille.

- *Ne requiert pas le génotypage des parents*. Certains logiciels qui implantent le test de déséquilibre de transmission (*transmission disequilibrium test*, TDT) requièrent que les parents soient génotypés, alors que beaucoup de parents de notre population d'étude ne sont pas génotypés. L'exclusion de ces sujets entraînerait une perte de puissance des tests. La méthode de rééchantillonnage proposée évite ce problème en ne transmettant pas aux enfants des génotypes parentaux réels mais plutôt en assignant d'abord aux fondateurs des haplotypes simulés qui respectent les principales caractéristiques des génotypes originaux puis en transmettant ces haplotypes aux enfants.
- *Respecte les fréquences alléliques originales*. Lors de la simulation des haplotypes parentaux, nous respectons les fréquences alléliques originales car la puissance d'un test d'association dépend ce cette fréquence: la puissance est très faible pour une fréquence proche de 0 (Wacholder et al., 2004). Ainsi, si la fréquence simulée est plus élevée que la fréquence originale, la correction est conservatrice (les simulations produisent plus facilement de bonnes valeurs p que les tests originaux). À l'inverse, si la fréquence simulée est plus faible que la fréquence observée, la correction est libérale. En respectant les fréquences originales, la correction est fiable. De plus, il n'est pas nécessaire d'établir un seuil arbitraire en deça duquel le test n'est pas effectué, contrairement à d'autres méthodes de correction.
- *Respecte le déséquilibre de liaison original*. Le déséquilibre de liaison réel est

estimé à partir des haplotypes (reconstruits) de nos sujets. Les haplotypes des fondateurs sont simulés en respectant le LD original ce qui permet à notre méthode d'être fiable. Elle serait conservatrice si elle supposait que le LD était nul entre tous les marqueurs mais serait libérale si elle « ajoutait » du LD inexistant. Lorsque deux SNPs sont corrélés (i.e. en LD), leurs résultats d'association sont également corrélés. À l'extrême, lorsque deux SNPs sont en parfait LD ($r^2 = 1$), ce ne sont pas deux tests statistiques différents qui sont effectués, mais bien le même test deux fois: il faut donc corriger pour un test et non pour deux. Les simulations permettent de tenir compte de tout le spectre de LD possible, sans fixer de seuil arbitraire.

- *S'applique à tous les tests d'association génomique familiale*. Notre méthode est générale et n'est pas basée sur un test d'association génomique familiale particulier. Elle peut donc être utilisée pour corriger les valeurs p de tous les tests d'association génomique familiale. En comparaison, certaines méthodes sont propres à un test d'association particulier, par exemple le TDT.
- *Respecte l'héritabilité des phénotypes*. Nous avons montré que le fait de ne pas respecter l'héritabilité des phénotypes pouvait produire des tests libéraux ou conservateurs. La méthode proposée respecte l'héritabilité en gardant intacts les phénotypes et en simulant plutôt les génotypes.

Chapitre 6: Aspects de génie logiciel

Nous avons développé un logiciel qui implante la méthode proposée. Nous avons également conçu et implanté une base de données relationnelle afin de conserver les expériences et les résultats qui seront décrits au chapitre suivant. Les principales activités de génie logiciel que nous avons menées sont: la conception du logiciel, son implantation et sa validation.

Quatre critères ont guidé le développement du logiciel. Premièrement, le développement doit être rapide: ainsi, si les expériences montrent que la méthode proposée n'est pas valide ou peu puissante, on pourra proposer une autre méthode. Deuxièmement, l'exécution du logiciel doit être assez rapide afin d'obtenir des valeurs p empiriques précises dans un délai raisonnable. Troisièmement, les expériences doivent être facilement repérables et on doit pouvoir facilement comparer les résultats entre expériences et entre méthodes de correction pour tests multiples. Finalement, le logiciel doit être fiable afin que les chercheurs qui l'utilisent aient confiance dans les résultats rapportés. Les sections suivantes montrent comment nous avons tenté de satisfaire les quatre critères.

6.1 Développement rapide

Afin d'accélérer le développement du logiciel, nous avons cherché un langage de programmation qui possède un niveau d'abstraction élevé et une librairie riche. Nous avons ensuite cherché et intégré des librairies et des logiciels tiers pour implanter certaines des étapes de la méthode. Un problème important que nous avons eu à résoudre est celui de l'interopérabilité des différentes composantes, dû aux différences dans les langages de programmation et dans la façon de communiquer des composantes.

6.1.1 Programmation en Python

L'implantation de la correction pour tests multiples s'est faite en Python[3], un langage de haut niveau typé dynamiquement. Python possède plusieurs qualités qui le rendent très approprié pour le développement de notre logiciel.

D'abord, le langage offre une programmation dans un niveau d'abstraction élevé, rendant le code compact et facile à comprendre. Par exemple, Python possède un ramasse-miettes, supporte les paramètres par défaut et par mots-clefs et offre de façon native des structures de données complexes comme des listes, des ensembles et des tables de dispersement. La même syntaxe permet d'itérer aisément sur une liste, un ensemble ou sur les lignes d'un fichier.

Ensuite, Python possède une librairie très étendue, entre autres un moteur d'expressions régulières, une interface aux ressources du système (pour démarrer un processus, changer de répertoire, etc.) et des interfaces aux principaux systèmes de gestion de bases de données.

Ces qualités font de Python un langage très approprié pour intégrer des composantes tiers. Les sections suivantes décrivent les composantes qui ont été intégrées.

6.1.2 fastPHASE

Nous avons retenu le logiciel *fastPHASE*[4]de Scheet et Stephens (2006) pour phaser les génotypes observés (figure 5.1, page 36, étape 2). Ce logiciel prend en entrée les génotypes observés et produit en sortie la paire d'haplotypes la plus probable pour chaque sujet génotypé.

Le logiciel modélise le déséquilibre de liaison par un modèle de Markov caché (*hidden Markov model*, HMM). Les auteurs partent de l'observation que les haplotypes des individus ont tendance à se regrouper en partitions (*clusters*) sur de courtes régions (quelques milliers de bases) mais que ces partitions varient le long du chromosome à

3 http://www.python.org/
4 http://depts.washington.edu/ventures/UW_Technology/Express_Licenses/fastPHASE.php

cause de la recombinaison. Ils permettent donc aux haplotypes observés de changer de partition d'un locus à un autre. Le logiciel suppose que les marqueurs sont en équilibre de Hardy-Weinberg.

Cette méthode est capable de modéliser autant un changement progressif du LD qu'un changement abrupt. Aussi, les estimés faits à partir de plusieurs points de départ sont combinés, ce qui évite de produire des haplotypes à partir d'un optimum local.

D'autres logiciels et d'autres méthodes sont disponibles pour phaser les génotypes. *fastPHASE* a été choisi parce que, parmi les logiciels capables de phaser précisément plusieurs centaines de SNPs chez des centaines de sujets, il est le plus rapide. Scheet et Stephens (2006) rapportent avoir phasé plus de 3 000 individus à près de 300 000 SNPs en 4 jours (processeur Intel Xeon cadencé à 3 GHz, 8 Go de RAM).

Notons que *fastPHASE* a été conçu pour être utilisé dans des études de type cas-témoins. Son application à une étude familiale n'a pas été évaluée. Par contre, les logiciels qui phasent les génotypes dans des familles ne peuvent pas traiter de grandes familles ni un grand nombre de marqueurs.

6.1.3 HapSim

Les haplotypes que nous assignons aux fondateurs (figure 5.1, étape 5) sont générés par la librairie *HapSim*[5] (Montana, 2005) utilisable à partir du logiciel statistique *R*[6] (R Development Core Team, 2006).

HapSim modélise un haplotype comme une variable aléatoire multivariée qui possède des distributions marginales connues (i.e. les fréquences alléliques) et des coefficients de corrélation connus (i.e. le LD entre chaque paire de SNPs). Pour chaque fondateur, deux vecteurs aléatoires sont simulés de la distribution multivariée normale dont le vecteur de la moyenne est (0, ..., 0). Ces vecteurs réels sont ensuite transformés en vecteurs d'allèles selon des seuils basés sur les fréquences alléliques originales.

5 http://cran.r-project.org/src/contrib/Descriptions/hapsim.html
6 http://www.r-project.org/

Montana (2005) démontre expérimentalement la grande ressemblance entre les propriétés des données simulées et celles des données originales. Lorsque la taille de l'échantillon simulé dépasse la centaine d'haplotypes, la différence entre le LD des haplotypes simulés et celui des haplotypes fournis au logiciel devient très faible, alors que la différence entre les fréquences alléliques simulées et observées devient presque nulle.

La limite principale de cette approche est que seule la corrélation entre des paires de SNPs est modélisée; la corrélation d'ordre plus élevé, par exemple entre des triplets de SNPs, n'est pas modélisée. Aussi, *HapSim* ne traite que les marqueurs bi-alléliques. Notre implantation est donc présentement limitée à ces marqueurs.

6.1.4 Merlin

Le logiciel *Merlin*[7] (Abecasis, 2002) a été retenu pour effectuer le *gene-dropping* (figure 5.1, étape 6). À la méthode de *gene-dropping* originale, *Merlin* ajoute la possibilité de transmettre les allèles de plusieurs gènes. Le logiciel prend en entrée les fréquences alléliques, les patrons de génotypes manquants et les fréquences de recombinaison entre les marqueurs et produit en sortie deux haplotypes par individu. Le logiciel génère d'abord les haplotypes des fondateurs. Puis, à chaque marqueur, l'allèle transmis à un enfant dépend de la fréquence de recombinaison θ entre le marqueur précédent et le marqueur courant: la probabilité de choisir l'allèle du même haplotype est $1-\theta$ alors que la probabilité de choisir l'allèle de l'autre haplotype est θ.

Nous avons modifié *Merlin* afin qu'il assigne aux fondateurs les haplotypes que nous avons simulés plutôt que de les générer lui-même. Nous avons ajouté la possibilité de produire des fichiers dans un format directement importable par *FBAT*, afin d'éviter une étape de conversion de fichiers qui peut être coûteuse en terme de temps d'accès au disque. Aussi, nous avons ajouté l'option d'afficher les individus que *Merlin* considère comme des fondateurs.

---

7 http://www.sph.umich.edu/csg/abecasis/Merlin/

6.1.5 Intégration

L'intégration des logiciels et librairies tiers pose plusieurs problèmes. Premièrement, des quatre logiciels utilisés (*FBAT*, *fastPHASE*, *HapSim* et *Merlin*), le code source de seulement deux d'entre eux est disponible; seul un exécutable est fourni pour *FBAT* et *fastPHASE*. Les deux logiciels dont le code source est disponible ne sont pas implantés dans le même langage de programmation: *Merlin* est implanté en C++ et *HapSim* en R.

Nous avons donc intégré les logiciels par exécution de processus externes. Notre programme fournit les données aux logiciels de trois façons: par fichiers, par le flot d'entrée standard et par des paramètres de la ligne de commande. Notre programme lit les fichiers de résultats produits par les logiciels tiers et extrait les valeurs d'intérêt.

Un problème que nous avons dû surmonter est que les formats d'entrée et de sortie de deux logiciels (*FBAT* et *fastPHASE*) ne sont pas complètement spécifiés. Nous avons donc dû faire plusieurs essais pour tenter d'inférer les formats, ce qui a mené à des surprises. Par exemple, *FBAT* termine anormalement et sans explications (*segmentation fault*) lorsqu'un des paramètres qu'on lui passe (le nom du fichier de phénotypes) dépasse 255 caractères. Aussi, le format de sortie de *FBAT* varie selon que le test FBAT a été effectué ou non (un SNP n'est pas testé si trop peu de familles sont informatives à ce SNP) et selon le type de test (multi-allélique ou bi-allélique). L'analyse des résultats de *FBAT* s'en trouve compliquée. De plus, lorsque le logiciel lit sur l'entrée standard le signal de fin de fichier (EOF), il part en boucle infinie en émettant continuellement le même symbole, plutôt que de quitter comme c'est la norme sous Linux. Si la sortie de FBAT est redirigée vers un fichier de sortie, le disque se remplit très rapidement.

Il faut aussi comprendre les paramètres à passer aux logiciels et identifier les résultats pertinents. Par exemple, *fastPHASE* produit deux ensembles d'haplotypes, la seule différence résidant dans le type de taux d'erreur que chaque ensemble minimise. Quant à *FBAT*, on doit comprendre la différence entre les deux hypothèses nulles, la transformation phénotypique la plus appropriée à apporter, les modèles génétiques, l'effet du nombre minimum de familles informatives, etc.

6.1.6 Fichier de configuration

Plusieurs informations doivent être transmises à notre programme: le nom des phénotypes, les chemins d'accès des fichiers qui contiennent les valeurs des phénotypes, les régions chromosomiques à analyser, le type d'analyse d'association à exécuter et les paramètres à lui passer, etc. Nous avons décidé de fournir ces informations dans un fichier de configuration qui respecte la syntaxe Python. Notre programme importe ce fichier de la même manière qu'une librairie Python quelconque (énoncé import). L'analyse lexicale et syntaxique est faite par l'interpréteur Python. Ainsi, nous n'avons pas à programmer une librairie d'analyse de fichiers de configuration, ce qui demanderait du temps et risquerait d'introduire des erreurs. De plus, tous les programmes que nous avons développés se servent du même fichier de configuration, ce qui évite des incohérences potentielles. Un extrait de fichier de configuration est présenté à la figure 6.1.

```
import libp1

simid = 17          # identificateur de l'expérience (tableau 6.1)
chr2orders = {      # région(s) chromosomique(s) à analyser
    '1' : [(3926, 3984)]}
chr2bounds = {}     # séparation des grands blocs (section 6.2.3)
N = 10000           # nombre de permutations
C = 50              # nombre de permutations par tâche
phenotypes = ['skinfold_body_fat',]  # phénotype(s) à analyser
ph2p = {            # meilleure valeur p observée à chaque phénotype
    'skinfold_body_fat' : 0.003661}
genetic_test = libp1.FbatTest(
    modes=['m',],   # modes FBAT ('m', 'b')
    models=['a',],  # modèles génétiques ('a', 'd', 'g', 'r')
    e_option=1,     # tester l'hypothèse nulle H02?
    rm_fbat_res=1,  # effacer les fichiers de résultats FBAT?
    minsize=5,      # nombre minimum de familles informatives
    nbest=1)        # nombre de meilleures valeurs p à stocker
# options de la base de données de simulations
sim_dbuser = 'brunelle'
sim_dbhost = '10.52.47.170'
sim_dbport = 5432
sim_dbname = 'simfbat'
```

Figure 6.1: Exemple de fichier de configuration

6.2 Exécution rapide

La méthode proposée est exigeante en terme de temps de calcul. La complexité de la méthode est cubique par rapport au nombre de marqueurs étudiés simultanément (section 6.4.1). Afin d'accélérer l'exécution des expériences, nous avons opté pour un calcul distribué et nous avons scindé les chromosomes en grands blocs.

Pour permettre un calcul distribué, nous avons divisé les expériences en tâches indépendantes qui peuvent être exécutées simultanément. La communication se fait via une base de données relationnelle.

Nous avons accès à huit serveurs de calcul, équipés de deux processeurs Intel P4 Xeon cadencés à 2,4 GHz et de 1,5 Go de mémoire vive. Nous avons aussi accès à un serveur de base de données exécutant le système de gestion de bases de données (SGDB) relationnel *PostgreSQL*. Tous les serveurs roulent un noyau Linux 2.4.

6.2.1 Division en tâches indépendantes

Les N simulations effectuées sont complètement indépendantes les unes des autres. De plus, les chromosomes sont indépendants les uns des autres car le déséquilibre de liaison et la liaison se limitent à des régions rapprochées d'un même chromosome. Il y a donc un parallélisme de haut niveau à exploiter.

Nous avons divisé le travail en tâches. Une tâche est définie comme un tuple (simulations, chromosome, région) qui représente tous les calculs devant être effectués sur une région d'un chromosome pour un certain nombre de simulations. En d'autres termes, une tâche peut comprendre plusieurs simulations et tous les phénotypes d'une expérience sont traités dans la même tâche. Ces deux éléments améliorent la performance. Premièrement, lorsque le nombre de tests à effectuer au cours d'une simulation est faible, il est avantageux d'augmenter le nombre de simulations par tâche afin de limiter le temps de communication entre le noeud de calcul et la base de données. Deuxièmement, l'étape la plus gourmande en temps de calcul (la simulation

des haplotypes) ne dépend pas des phénotypes. En traitant tous les phénotypes dans une même tâche, on simule les haplotypes une seule fois puis on teste plusieurs phénotypes.

6.2.2 Approche client-serveur

Nous avons adopté un modèle client-serveur, dans lequel le logiciel de correction pour tests multiples demande à un maître une tâche à exécuter. Un système de gestion de bases de données (SGBD) a été choisi pour jouer le rôle de maître afin de minimiser la quantité de logiciel à développer, d'obtenir une fiabilité et une sécurité élevées et de gérer la concurrence.

Avant de débuter une expérience, on la divise en tâches qui sont insérées dans la base de données. D'autres informations sont associées à la tâche: une graine aléatoire pour *HapSim* et une autre pour *Merlin* ainsi que les temps de début et de fin de la tâche, afin de déterminer ultérieurement le temps de calcul et le temps écoulé d'une expérience.

Lorsqu'un serveur démarre ou lorsqu'il vient de terminer une tâche, il fait une requête à la base de données afin d'obtenir une tâche; si une tâche est disponible, il met à jour l'attribut de temps initial de la tâche. Sinon, le serveur arrête. Ces opérations se font dans le cadre d'un énoncé SELECT FOR UPDATE afin d'éviter des problèmes de concurrence. Le serveur insère, pour chaque simulation et chaque phénotype, le meilleur résultat FBAT dans la base de données. Lorsque la tâche est terminée, le serveur met à jour l'attribut du temps final de la tâche.

Notons que le serveur de base de données représente un point unique de défaillance: s'il tombe en panne, les clients ne peuvent plus travailler.

6.2.3 Division d'un chromosome en grands blocs

En pratique, les serveurs de calcul manquent de mémoire lorsqu'ils traitent environ 1000 SNPs. Puisque pour certains chromosomes nous avons de l'information sur plus de 4500 SNPs, nous avons divisé les chromosomes en grands blocs. Puisque chaque bloc est traité indépendamment, le LD entre deux blocs ne peut être pris en compte: nous avons

donc intérêt de trouver des endroits qui montrent peu de LD. Les blocs ont été délimités en fonction de deux critères: d'une part, le taux de recombinaison doit être élevé entre deux blocs (ce qui implique en général un faible LD); d'autre part, un bloc doit contenir au maximum 600 SNPs. Cette division a produit entre un et 11 blocs par chromosomes (135 blocs pour les 22 autosomes, soit environ 400 SNPs par bloc en moyenne).

6.3 Conservation des expériences

En plus de permettre un calcul distribué, la base de données sert également à gérer les expériences et les résultats. Cette gestion vise deux buts: d'abord de pouvoir facilement répéter une expérience si le besoin s'en fait sentir; ensuite de pourvoir facilement comparer les expériences entre elles ou les méthodes de correction entre elles. Nous avons conçu quelques tables, présentées au tableau 6.1. La gestion des expériences se fait par les tables *simulation* et *task*. De façon pratique, le champ *description* de la table *simulation* pointe au répertoire du test *FBAT* original; ce répertoire contient les fichiers de configuration nécessaires pour refaire l'expérience. Les expériences décrites au chapitre 7 correspondent chacune à un tuple de la table *simulation*.

6.4 Implantation

La figure 6.2 présente les étapes pour calculer la valeur p empirique de la meilleure association d'une expérience. Avant de débuter ces étapes, l'usager doit avoir préparé le fichier de configuration, décrit à la section 6.1.6.

D'abord, les expériences sont divisées en tâches qui sont insérées dans la base de données (ligne 1). Ensuite, les génotypes sont phasés par le logiciel *fastPHASE* (ligne 2). Ces étapes sont exécutées une seule fois par expérience. Les étapes 3 à 13 sont exécutées un grand nombre de fois et sont parallélisables. Tant qu'une tâche est disponible (lignes 4 et 5), un noeud de calcul calcule le déséquilibre de liaison (ligne 6), puis, pour chaque simulation, simule les haplotypes des fondateurs par HapSim (ligne 8), transmet les allèles des fondateurs à leurs descendants (ligne 9), calcule les

associations entre tous les SNPs et tous les phénotypes (ligne 11) et insère le meilleur résultat FBAT de chaque phénotype dans la base de données (ligne 12). Une fois les simulations terminées, le noeud de calcul indique dans la base de données que la tâche est terminée. Finalement, une fois toutes les tâches terminées, il est possible de calculer la valeur p empirique (ligne 14).

Tableau 6.1: Tables de la base de données de gestion des expériences

Table	***Champ***	***Description***
simulation	id	Identificateur unique de l'expérience
	description	Description textuelle de l'expérience
task	simid	Identificateur de l'expérience (id de simulation)
	chromosome	Chromosome (1-22)
	base_replicate	Indice de la 1ère simulation à effectuer
	replicates	Nombre de simulations à effectuer
	min_order	Indice du premier SNP à traiter
	max_order	Indice du dernier SNP à traiter
	merlin_seed	1ère graine aléatoire à fournir à *Merlin*
	hapsim_seed	Graine aléatoire à fournir à *HapSim*
	server	Nom du serveur sur lequel la tâche a été démarrée
	begtime	Temps du début de la tâche
	endtime	Temps de complétion de la tâche
	pid	Identificateur du processus qui exécute la tâche
sim_result	simid	Identificateur de l'expérience (id de simulation)
	chromosome	Chromosome (1-22)
	mkname	Nom du marqueur
	phenotype_id	Identificateur du phénotype
	replicate	Réplicat
	p	Valeur p
bounds	simid	Identificateur de l'expérience (id de simulation)
	chromosome	Chromosome (1-22)
	bound	Indice du SNP qui termine un grand bloc

1. Diviser l'expérience en tâches et les insérer dans la base de données
2. Phaser les génotypes (*fastPHASE*)
3. RÉPÉTER indéfiniment
4. Obtenir de la base de données une tâche *t*
5. SI il n'y a plus de tâche à effectuer ALORS quitter
6. Calculer le LD
7. POUR CHAQUE simulation *s*
8. Simuler les haplotypes des fondateurs (*HapSim*)
9. Transmettre les allèles des fondateurs à leurs descendants (*Merlin*)
10. POUR CHAQUE phénotype *p*
11. Calculer les associations (*FBAT*)
12. Insérer le meilleur résultat de *FBAT* dans la base de données
13. Marquer la tâche *t* comme étant terminée
14. Calculer la valeur p empirique

Figure 6.2: Exécution d'une expérience

Les fonctionnalités sont implantées par cinq programmes Python, énumérés au tableau 6.2. La correspondance entre le programme Python et les étapes de la figure 6.2 est indiquée à la fin du champ « responsabilités ».

Tableau 6.2: Logiciels utilisés pour la correction

Logiciel	***Responsabilités***
p1-fill.py	Divise l'expérience en tâches et génère les énoncés SQL pour insérer les tâches dans la base de données (ligne 1)
p1-init.py	Exécute *fastPHASE*, extrait les patrons de génotypes manquants des données originales et divise les chromosomes en blocs (ligne 2)
assoc-simulation.py	Demande une tâche, exécute *HapSim*, *Merlin* et *FBAT* puis insère les résultats dans la base de données (lignes 3 à 13)
p1-remote-starter.py	Démarre et arrête *assoc-simulation.py* à distance
sim-analysis.py	Extrait les résultats et rapporte les valeurs p corrigées (ligne 14)

La plupart des fonctions et classes ont été regroupées au sein d'une librairie nommée *libp1.py*. Les classes de cette librairie sont présentées à la figure 6.3.

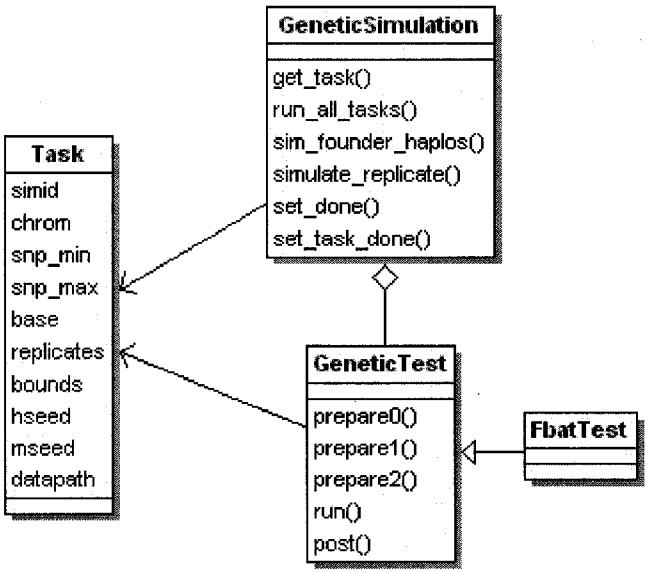

Figure 6.3: Diagramme des principales classes

La classe *GeneticSimulation* est responsable d'exécuter des tâches; sa méthode principale est *run_all_tasks*.

La classe *Task* représente une tâche. Ses principaux attributs sont *simid* qui identifie une expérience, *chrom*, *snp_min* et *snp_max* qui identifient une région chromosomique à analyser et *base* et *replicates* qui indiquent les simulations à effectuer.

La classe *GeneticTest* représente le test génétique de base. C'est une classe abstraite de laquelle dérive *FbatTest*. D'autres tests d'association peuvent être ajoutés au logiciel en dérivant de *GeneticTest*. Trois méthodes sont appelées avant que le test ne soit exécuté: *prepare0* est appelée avant que la première tâche ne soit obtenue (c'est le bon endroit pour vérifier que les librairies et exécutables nécessaires au test génétique existent), *prepare1* avant la première simulation de chaque tâche et *prepare2* avant chaque simulation. La méthode *run* exécute une simulation sur tous les marqueurs et un phénotype. Finalement, *post* est appelée après chaque simulation. Un exemple de séquence d'appels est présenté à la figure 6.4.

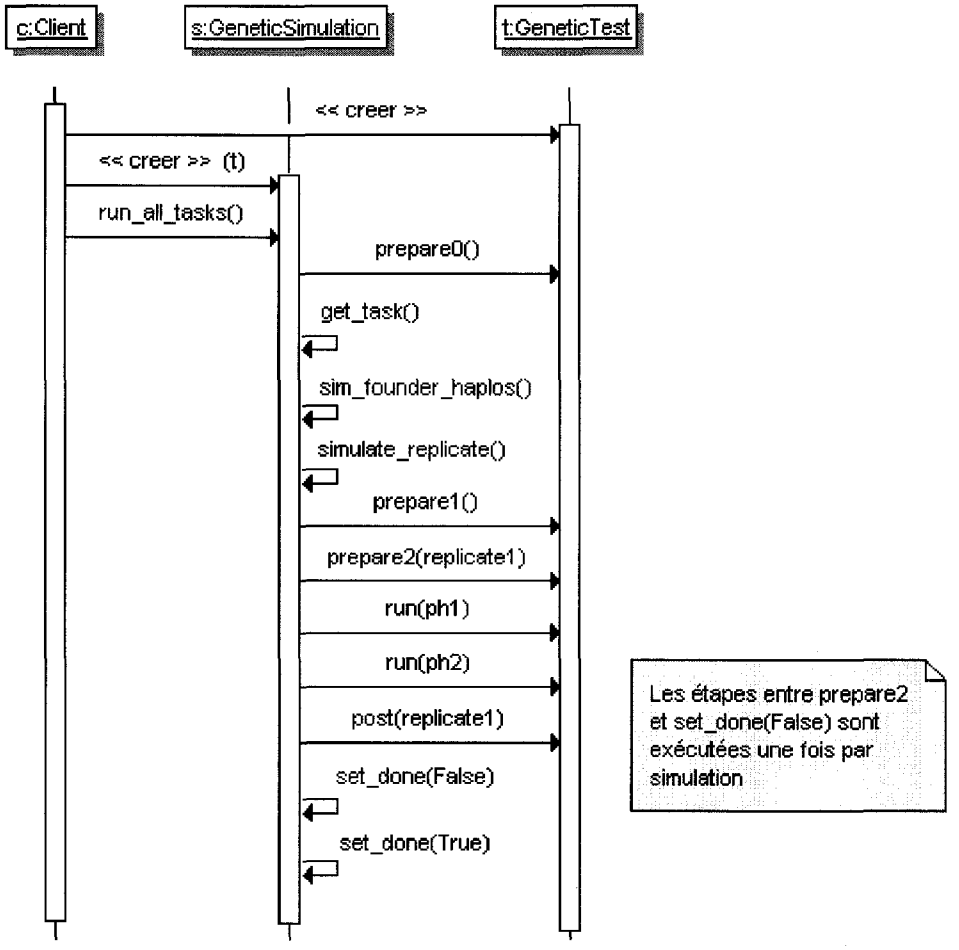

Figure 6.4: Diagramme de séquence d'une simulation

6.4.1 Complexité algorithmique

Examinons comment le temps d'exécution de l'implantation présentée à la figure 6.2 varie en fonction des paramètres d'entrée. Dans ce qui suit, notons N le nombre de simulations, n le nombre de sujets, f le nombre de fondateurs, m le nombre maximal de marqueurs dans un grand bloc, b le nombre de grands blocs, M le nombre de marqueurs au total, P le nombre de phénotypes et K le nombre de partitions utilisées par *fastPHASE*.

La complexité de *fastPHASE* (ligne 1) est $O(nK^2M)$ (Scheet & Stephens, 2006). En

particulier, l'algorithme est linéaire par rapport au nombre de sujets et au nombre de marqueurs. En pratique, K est une constante choisie entre 5 et 20.

Le calcul du LD (ligne 6) est de complexité $O(m^2 n)$ (voir équations 2.2 et 2.4, section 2.3).

HapSim (ligne 8) fait appel à la fonction R *mvrnorm* pour calculer les valeurs propres et les vecteurs propres de la matrice de corrélation entre les marqueurs et pour générer des vecteurs aléatoires. Le calcul des valeurs propres et vecteurs propres se fait par la librairie Fortran *EISPACK* (Bowdler et al., 1968; Martin et al., 1968) dont la complexité est $O(m^3)$. La génération des vecteurs aléatoires se fait par deux multiplications matricielles: la première multiplication entre deux matrices de m rangées et m colonnes et la seconde entre une matrice de m rangées et m colonnes et une autre de m rangées et n colonnes. La multiplication matricielle est implantée par la fonction *DGEMM* de la librairie *BLAS*. La complexité de *DGEMM* pour une matrice de A rangées et B colonnes et une autre de B rangées et C colonnes est $O(ABC)$ La complexité des deux multiplications est $O(m^3+m^2 n)$. La complexité de *HapSim* est donc $O(m^3+m^2 n)$.

Le *gene-dropping* (ligne 9) est très efficace, autant en terme de temps de calcul que de mémoire: au total, $2(n-f)m$ transmissions sont effectuées. La complexité est $O(mn)$, donc linéaire par rapport au nombre de marqueurs et au nombre de sujets étudiés.

La complexité de FBAT (ligne 11) est de $O(mn)$ (voir section 3.1).

Lorsqu'une tâche couvre toutes les simulations pour un bloc, la ligne 6 est exécutée b fois, les lignes 8 et 9 bN fois et la ligne 11 bNP fois. La complexité de l'algorithme en son entier est donc $O(nK^2 M+bm^3 N+bm^2 nN+bmnNP)$.

En pratique, le temps de calcul dû aux second et troisième termes est de loin supérieur aux premier et quatrième termes, de sorte que la complexité est $O(bm^3 N+bm^2 nN)$.

Pour ce qui est de la complexité en terme de mémoire vive, la génération des haplotypes demande de conserver une matrice de taille m par m et une autre de taille m par n. La

complexité en terme de mémoire est donc $O(m^2+mn)$.

6.4.2 Limites

Notre approche est présentement limitée aux marqueurs bi-alléliques du fait de l'utilisation du logiciel *HapSim*. En pratique cette limitation a peu de conséquences puisque la plupart des études d'association se font sur des SNPs.

Le serveur de base de données représente un point unique de défaillance: s'il tombe en panne, les clients ne peuvent plus travailler. Toutefois, aucune information n'est perdue.

6.5 Fiabilité

Le logiciel développé repose en partie sur des logiciels et des librairies externes. Nous avons donc validé les logiciels externes et les modifications que nous avons apportées à ces logiciels.

Pressman (1997) définit la validation comme étant l'ensemble des activités visant à s'assurer que le logiciel répond aux requis du client. Voici les principales validations qui ont été effectuées:

1. s'assurer que les haplotypes simulés par *HapSim* respectent les fréquences alléliques et le déséquilibre de liaison des génotypes originaux;
2. s'assurer que les génotypes simulés par la version de *Merlin* modifiée respectent la transmission mendélienne, les fréquences alléliques originales, les fractions de recombinaison, le LD et les patrons de génotypes manquants;
3. s'assurer que la méthode est fiable (i.e. ni libérale, ni conservatrice).

Dans les deux premiers cas, il s'agit de valider un logiciel développé par un tiers, alors que dans le dernier cas, il s'agit de valider la méthode comme un tout.

Avant d'examiner les validations effectuées, une menace importante à la validité des résultats est la qualité des générateurs de nombres pseudo-aléatoires utilisés. Nous

examinons donc dans la prochaine section les qualités d'un bon générateur et présentons les générateurs utilisés.

6.5.1 Génération de nombres pseudo-aléatoires

Trois des logiciels utilisés (*HapSim*, *Merlin* et *fastPHASE*) utilisent un générateur de nombres pseudo-aléatoires (*pseudo-random number* generator, PRNG). Westfall et Young (1993, p. 18) insistent sur l'importance d'utiliser un PRNG qui se rapproche autant que possible d'un générateur de nombres aléatoires idéal afin que les simulations ne soient pas biaisées. Hellekalek (1998) a listé les propriétés que les nombres générés par un bon PRNG devaient posséder. Les principales sont: une distribution uniforme, une absence de corrélation, une reproductibilité et une période très grande par rapport à la taille de la séquence générée. Nous examinons dans cette section la qualité des PRNG utilisés.

Examinons d'abord *HapSim* et *Merlin.* L'environnement statistique *R* (2.1.1) et le logiciel *Merlin* (1.0.1) utilisent le PRNG Mersenne Twister récemment développé par Matsumoto et Nishimura (1998). Il possède une période de $2^{19937} - 1$ et une équidistribution dans 623 dimensions avec une précision de 32 bits ou encore dans 19937 dimensions avec une précision de 1 bit. Rappelons que la période d'un PRNG est la taille de la plus grande séquence qui peut être générée avant que la séquence ne se répète. Quant à elle, l'équidistribution de dimension k peut être définie de plusieurs façons. Une définition basée sur la cryptographie est la suivante: la connaissance des m premiers nombres ne donne strictement aucune information sur le prochain nombre, si $m < k$ (Matsumoto & Nishimura, 1998).

Le PRNG a passé avec succès la batterie de tests Diehard de Marsaglia et la batterie de tests Load et Ultimate Load de Wegenkittl (Matsumoto & Nishimura, 1998). De plus, le générateur n'utilise que 624 entiers de mémoire pour stocker son état et sa vitesse est comparable aux PRNG couramment utilisés. Par exemple, il est pratiquement aussi rapide que la fonction rand de la librairie C mais sa période de $2^{19937} - 1$ se compare

avantageusement à la période de 2^{31} de rand. En conclusion, le PRNG Mersenne Twister possède d'excellentes propriétés.

Nous associons à chaque tâche deux graines d'initialisation du PRNG, une pour *HapSim*, l'autre pour *Merlin*: nous voulons que toutes les simulations de génotypes démarrent d'une graine différente, sinon elles produiront des résultats identiques. Les tâches peuvent ensuite être exécutées sur un seul ou plusieurs ordinateurs. Cette approche de construction d'un PRNG parallèle pour simulations a été proposée par Hellekalek (1998) sous le nom *splitting technique*. C'est l'approche préconisée par Entacher et al. (1998).

Quant à lui, *fastPHASE* appelle la function drand48 de la librairie C stdlib.h. Cette fonction implante un générateur congruentiel linéaire (*linear congruential generator*, LCG) de période 2^{48}. Un LCG implante la fonction de récurrence:

$$X_{n+1} = (aX_n + c) \bmod m$$

La qualité du LCG dépend des paramètres a, c et m. Les LCG souffrent de plusieurs problèmes: la période est restreinte, l'équidistribution est faible et les bits les moins significatifs sont moins aléatoires que les bits les plus significatifs. La fonction est appelée à deux endroits par *fastPHASE*: pour initialiser les paramètres du modèle de Markov et pour estimer les haplotypes les plus probables des individus. Selon les auteurs, la faiblesse du PRNG devrait avoir peu d'impact sur la qualité de la solution (Scheet & Stephens, communication personnelle).

6.5.2 Validation des haplotypes simulés par *HapSim*

Nous avons exécuté *HapSim* sur des haplotypes constitués de 49 SNPs du chromosome 2 afin de générer 100 groupes de 100 haplotypes chacun. Le but est de s'assurer que les fréquences alléliques et le LD des haplotypes simulés ressemblent à ceux des haplotypes fournis en entrée. Nous avons utilisé les données qui seront décrites au chapitre 7. Nous analysons trois des 100 groupes (groupes 75, 76 et 77).

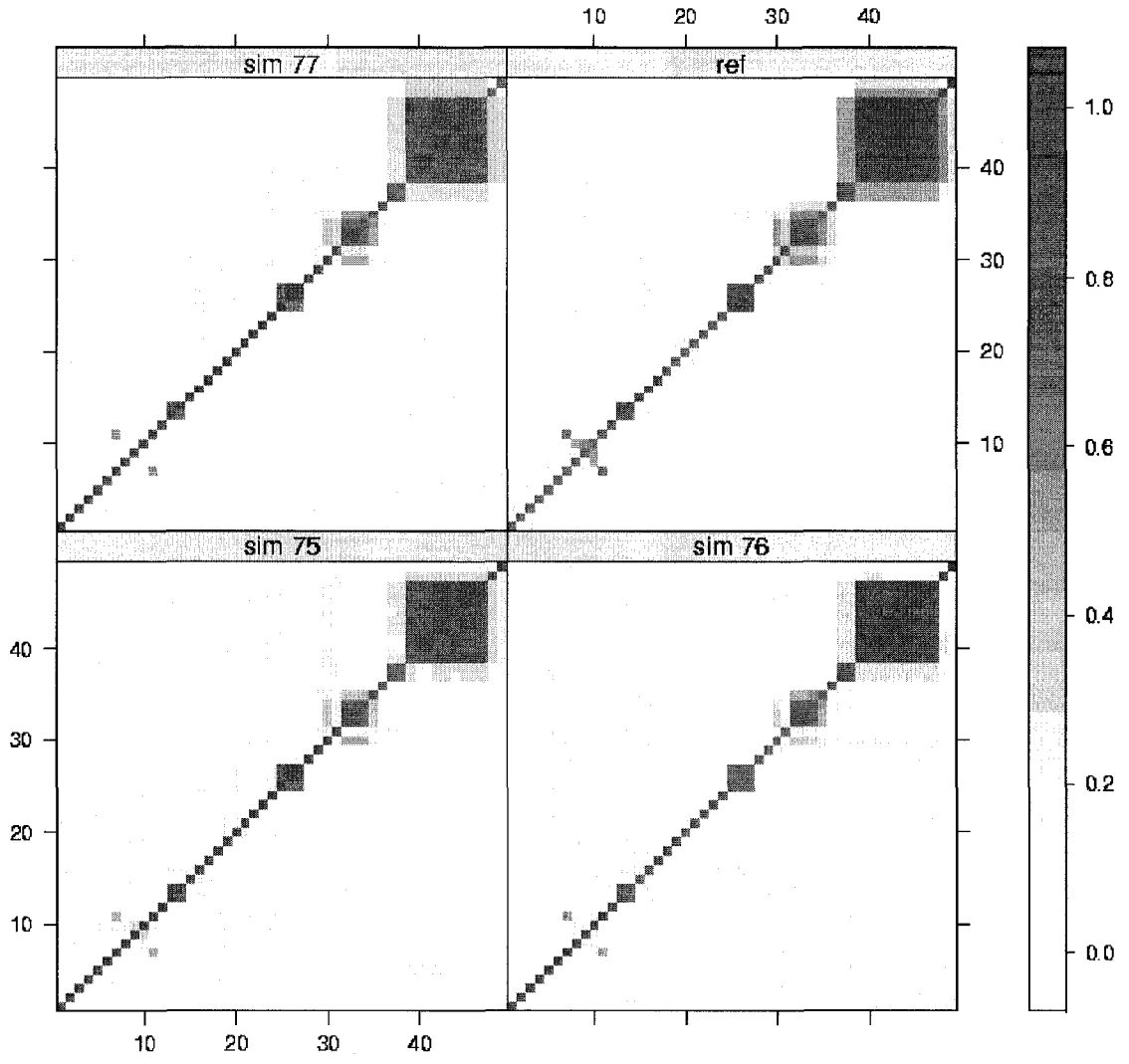

Figure 6.5: Le LD des haplotypes simulés par HapSim respecte le LD original

Le LD original est étiquetté « ref » (en haut à droite). Les cases « sim 75 », « sim 76 » et « sim 77 » correspondent à trois des 100 groupes générés. La métrique r^2 est utilisée. Elle varie de 0 (aucun LD, montré en pâle) à 1 (LD total, montré en foncé). Dans une case, l'abscisse et l'ordonnée représentent les SNPs en ordre croissant de position sur le chromosome.

La corrélation entre les fréquences alléliques originales et simulées calculée par le coefficient de corrélation de Pearson varie entre 98,6% et 99,1% pour les trois groupes. La plus grande différence pour un SNP donné est de 10% (fréquence simulée de 21%

contre fréquence originale de 31%). Lorsqu'on augmente le nombre d'haplotypes générés à 1000, la précision augmente: la corrélation varie entre 99,8% et 99,9% et la plus grande différence est de 3,8%. On conclut que les fréquences alléliques observées et simulées sont très semblables.

La figure 6.5 compare le LD entre les haplotypes observés et les haplotypes simulés, pour les trois mêmes simulations. On constate une très grande concordance entre le LD observé et le LD simulé. En particulier, la région de très grand LD (SNPs 38 à 48) est bien conservée. On peut donc s'attendre à ce que nos tests ne soient pas indûment conservateurs à cause des SNPs qui sont en LD.

La validation de la librairie *HapSim* nous a permis de découvrir une erreur dans la fonction *haplosim*. Lorsqu'une option qui vaut par défaut FAUX est mise à VRAI, une boucle de la fonction exécute toujours une seule itération, plutôt qu'un certain nombre. L'erreur a été rapportée au développeur de *HapSim*.

6.5.3 Validation de Merlin

Le tableau 6.3 présente les tests que nous avons menés pour valider la version modifiée de *Merlin*.

Nous voulons d'abord nous assurer que les allèles simulés qu'on retrouve chez les enfants proviennent de leurs parents. Une erreur mendélienne indiquerait une erreur dans le mécanisme de *gene-dropping* de *Merlin*. On s'assure donc, après l'exécution de *Merlin*, pour chaque enfant et chaque marqueur, que l'allèle identifié comme provenant d'un parent se retrouve bel et bien dans le génotype de ce parent. L'examen des chromosomes 1 et 22 et de génotypes simulés (voir prochain paragraphe) n'a révélé aucune erreur.

Pour les distances génétiques, nous avons simulé 1000 familles nucléaires comprenant deux parents et huit enfants et une carte de 100 marqueurs, espacés les uns des autres d'une distance comprise entre 0,1 et 5 cM tirée aléatoirement selon une distribution uniforme. Un grand nombre de familles a été simulé car le nombre prévu de

recombinaisons entre deux marqueurs est relativement faible. La corrélation de Pearson entre les distances génétiques originales et les fréquences de recombinaison simulées a été calculée pour les 99 paires de marqueurs adjacents. Cette corrélation est de 98,7%, ce qui indique que les génotypes simulés par *Merlin* respectent de très près les distances génétiques qui lui sont fournies.

Tableau 6.3: Tests de Merlin

Propriété testée	***Mesure***
Transmission mendélienne	Nombre d'allèles d'enfants qui ne proviennent pas d'un parent (i.e. nombre d'erreurs mendéliennes)
Distances génétiques	Corrélation de Pearson entre distances génétiques originales et simulés
Patrons de génotypes manquants	Nombre de génotypes manquants pour lesquels un génotype non-manquant a été simulé et nombre de génotypes non-manquants pour lesquels un génotype manquant a été simulé
Haplotypes des fondateurs	Pour les fondateurs, nombre de génotypes simulés différents des génotypes originaux
LD	Corrélation de Pearson entre LD original et simulé
Fréquences alléliques	Corrélation de Pearson entre fréquences alléliques originales et simulées

Pour les patrons de génotypes manquants, nous avons utilisé les données originales des chromosomes 1 et 22 et avons effectué une simulation. Tous les génotypes originaux qui étaient manquants le sont également dans les données simulées, et à l'inverse aucun génotype original présent n'est manquant dans les données simulées. *Merlin* respecte donc les patrons de génotypes manquants.

Nous voulons aussi que *Merlin* respecte les haplotypes assignés aux fondateurs, c'est-à-dire qu'il ne modifie pas leurs génotypes. La comparaison montre que tous les génotypes des fondateurs du chromosome 22 sont identiques avant et après une simulation, sous réserve que le génotype du fondateur était connu dans les données originales (rappelons que nous assignons des haplotypes à tous les fondateurs afin qu'ils les transmettent à

leurs descendants lors du gene-dropping, mais que pour le test *FBAT* seuls les fondateurs génotypés doivent avoir des génotypes simulés, les autres doivent avoir des génotypes manquants).

Deux mesures du LD ont été comparées: D' et r^2. Les SNPs du chromosome 22 ont été comparés. Ce chromosome a été choisi parce qu'il est très court. La comparaison se fait entre les haplotypes produits par *HapSim* (fournis à *Merlin*) et ceux que *Merlin* produit. Le LD est calculé entre un SNP et les cinq SNPs qui le suivent. La corrélation de Pearson est de 90,0% pour D' et de 99,9% pour r^2. Cette différence peut s'expliquer par le fait que D' est très instable pour de faibles fréquences alléliques. Nous considérons que *Merlin* respecte le LD des données d'entrée.

Finalement, la comparaison entre les fréquences alléliques originales et simulées s'est faite sur les chromosomes 1 et 22. La corrélation de Pearson est de 97,9% pour le chromosome 1 et de 98,0% pour le chromosome 22, ce qui indique que les fréquences alléliques sont respectées.

En somme, les génotypes simulés par *Merlin* respectent les propriétés fournies en entrée. La prochaine section examine la méthode en son entier.

6.5.4 Validation du contrôle de FWER

La méthode proposée a été validée par simulations. Les génotypes observés du chromosome 22 ont été utilisés et des phénotypes ont été simulés sans tenir compte des génotypes, mais selon une certaine héritabilité. Ainsi, il n'y a aucune association entre les génotypes observés et les phénotypes simulés. Pour chaque simulation de phénotype, *FBAT* est exécuté, la meilleure association est conservée puis ajustée à l'aide de la présente méthode de correction pour tests multiples. Puisque l'hypothèse nulle conjointe est vraie (i.e. aucun SNP n'est associé au phénotype), nous nous attendons à ce que les valeurs p corrigées proviennent de la distribution uniforme entre 0 et 1 (section 4.1).

De façon plus formelle, la procédure suivante est exécutée (inspirée de Westfall & Young, 1993, pp. 38-39).

1. RÉPÉTER D fois
2. Simuler un phénotype dont 20% de la variance est expliquée par un effet génétique, 20% par un effet polygénique, 20% par un effet familial et 40% est aléatoire. Le phénotype est simulé sans tenir compte des génotypes.
3. Exécuter la méthode de correction pour tests multiples proposée (section 5.2) sur le phénotype simulé et les génotypes originaux, en exécutant N réplicats; conserver la meilleure valeur p corrigée.

Figure 6.6: Validation du contrôle de FWER

Le chromosome 22 a été choisi car c'est celui qui contient le moins de SNPs (seulement 328), ce qui nous permet de faire $D = 4000$ répétitions de $N = 1000$ simulations chacune. Cela représente tout de même 1,3 milliard de tests FBAT. Le choix de $D = 4N$ respecte une suggestion de Westfall et Young. La simulation du phénotype est réalisée grâce au logiciel *SIBSIM* (Franke et al., 2006), qui permet de spécifier le pourcentage de variance du phénotype qui est expliqué par différentes causes. La distribution des $D = 4000$ valeurs p corrigées est présentée à la figure 6.7. Chaque valeur p corrigée est mise en relation avec une valeur p attendue: lorsque n nombres sont tirés de façon indépendante d'une distribution uniforme entre 0 et 1, la valeur attendue du $i^{\text{ème}}$ plus petit nombre est $i/(n+1)$, pour i allant de 1 à n. Puisque les $D = 4000$ répétitions sont faites sous l'hypothèse nulle et que chaque répétition est indépendante des autres, les valeurs p corrigées par une méthode fiable devraient suivre cette distribution. On constate que les valeurs p corrigées par la méthode proposée suivent de très près les valeurs p attendues. D'ailleurs un test de Kolmogorov-Smirnov (hypothèse nulle: les valeurs p proviennent d'une distribution uniforme entre 0 et 1) rapporte une valeur p de 0,7: il n'y a aucune raison de croire que notre méthode n'est pas fiable. En particulier, le tableau 6.4 montre que la proportion de valeurs p plus petites ou égales à q est très proche de q pour de petites valeurs q. Par exemple, exactement 1% des valeurs p corrigées sont plus petites ou égales à 1% et 9,3% sont plus petites ou égales à 10%.

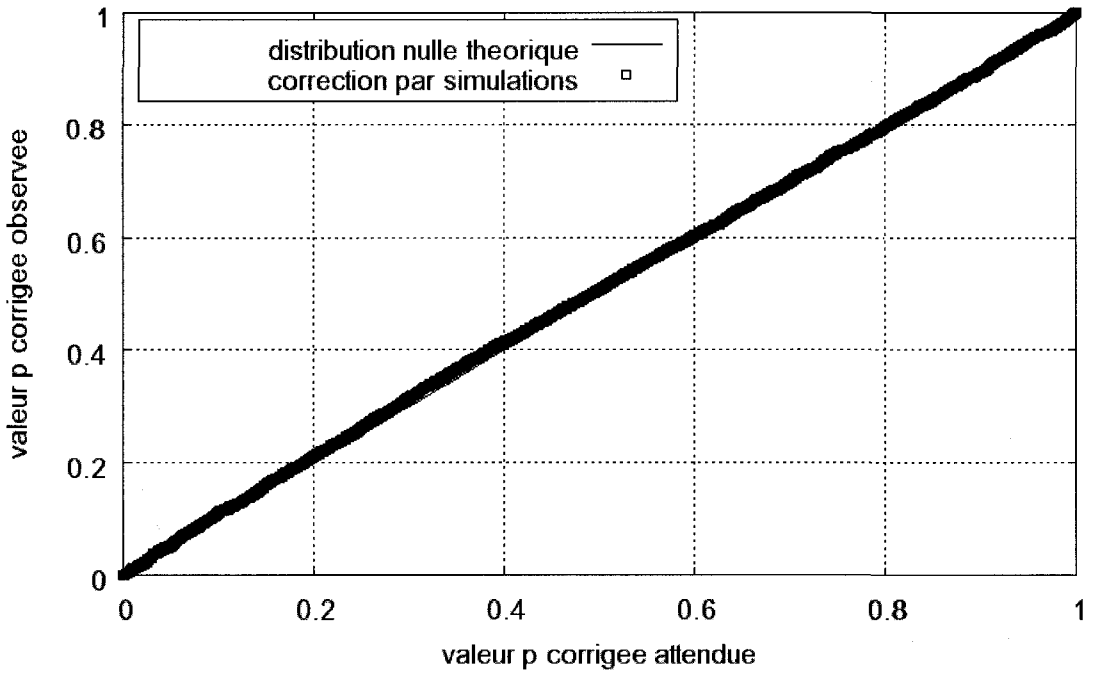

Figure 6.7: Distribution des valeurs p rapportées par la méthode proposée

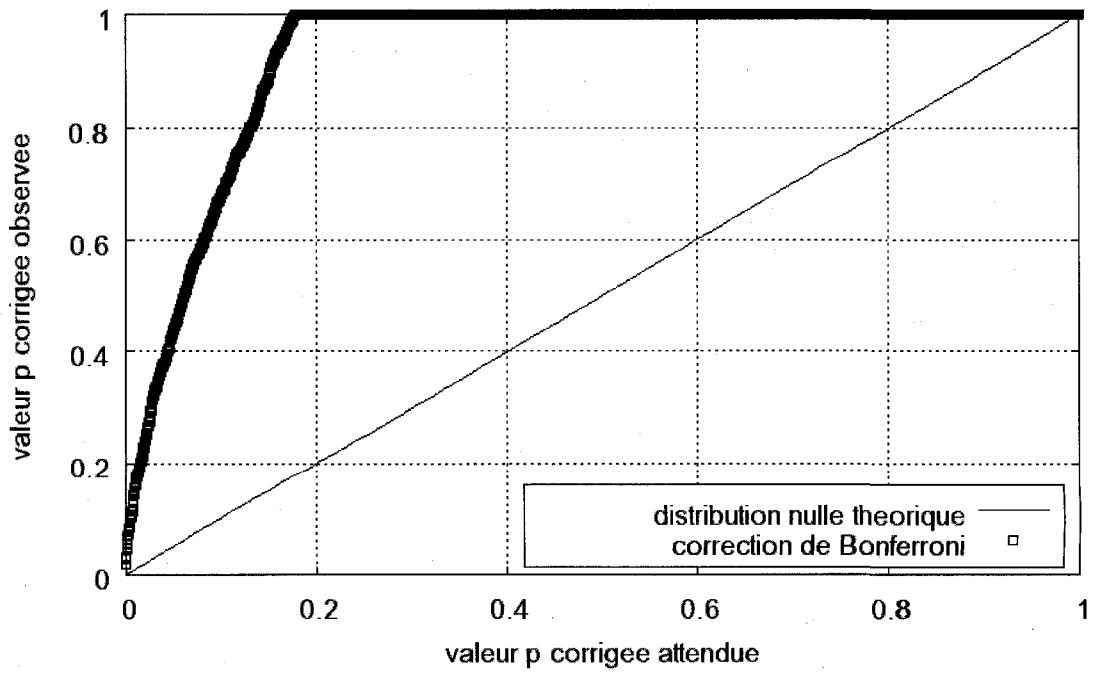

Figure 6.8: Distribution des valeurs p rapportées par la correction de Bonferroni

Tableau 6.4: Proportion de valeurs p inférieures ou égales à un seuil nominal

Seuil nominal	***Simulations***	***Bonferroni***
0,010	0,010	0,000
0,050	0,048	0,001
0,100	0,093	0,005

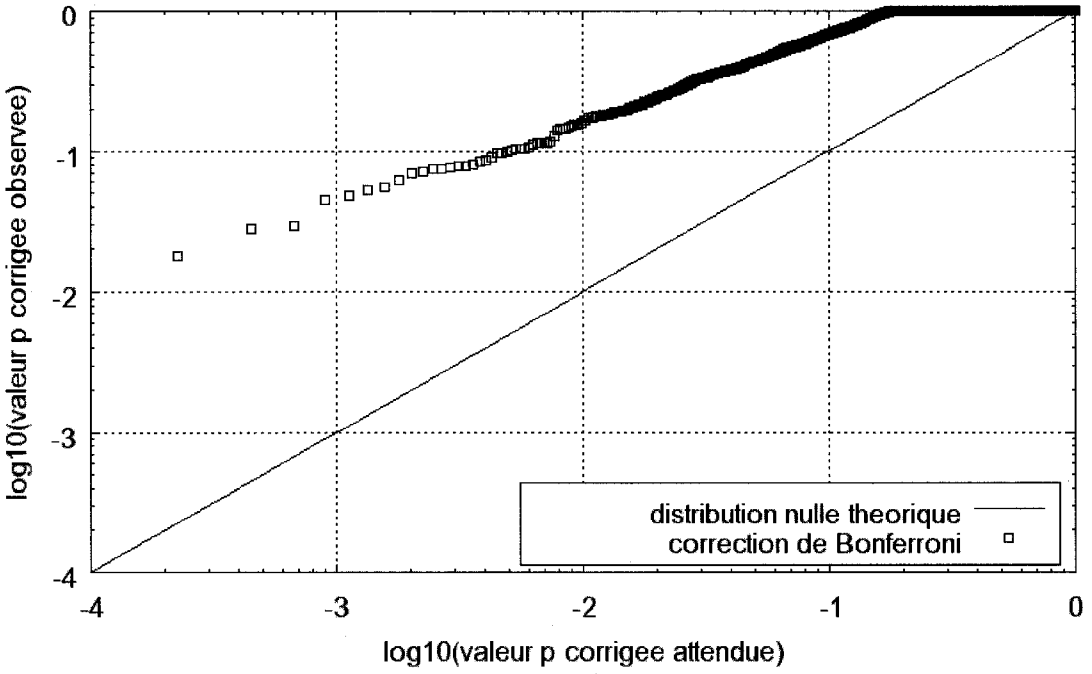

Figure 6.9: Échelle logarithmique afin de mettre l'emphase sur les faibles valeurs p

À l'opposé, la correction de Bonferroni est extrêmement conservatrice. Par exemple, seulement 0,5% des valeurs p corrigées sont plus petites ou égales à 10% (tableau 6.4). Les figures 6.8 et 6.9 montrent la distribution des valeurs p corrigées par Bonferroni. Les valeurs p sont très loin de la droite attendue. La figure 6.8 donne l'impression que les valeurs p rejoignent la droite attendue pour de faibles valeurs p mais une échelle logarithmique (figure 6.9) montre que ce n'est pas le cas: aucune valeur p corrigée n'est inférieure à 0,01 alors qu'on devrait obtenir quelques valeurs p corrigées inférieures à 0,001, même en l'absence d'association. Ces résultats sont très limités: ils ont été

produits sur une structure généalogique particulière, une distribution phénotypique particulière et des génotypes particuliers. Tout de même, ils nous amènent à penser que dans certaines situations la méthode proposée pourrait être beaucoup plus puissante que la correction de Bonferroni.

6.5.5 Sommaire des validations

Les validations que nous avons effectuées sont satisfaisantes: *HapSim* et *Merlin* respectent les propriétés des données génétiques qui leur sont fournies, alors que la méthode proposée contrôle adéquatement l'erreur de type I. Ces validations nous donnent confiance que les valeurs p rapportées par notre méthode sont fiables. Par contre, les validations ne sont en aucun cas exhaustives et n'impliquent pas que le logiciel ne contient pas de défauts (Pressman, 1997). En particulier, une seule validation a été effectuée sur la méthode en son entier. D'autres validations augmenteraient notre confiance ou permettraient de découvrir des défauts. Il serait intéressant de tester d'autres structures familiales, d'autres distributions de phénotypes et un nombre de marqueurs plus grand ou plus petit.

Chapitre 7: Expériences et résultats

Nous débutons cette section par une description de la population étudiée. Nous décrivons le phénotypage et le génotypage effectués sur cette population et les paramètres que nous devons spécifier pour exécuter les tests FBAT. Puis, nous décrivons huit expériences que nous avons menées afin de trouver des associations statistiquement significatives après correction pour tests multiples et de comparer la méthode proposée à la correction de Bonferroni dans plusieurs contextes. Nous présentons ensuite les résultats que nous avons obtenus: les valeurs p observées et les valeurs p corrigées par notre méthode et par la correction de Bonferroni. Nous commentons ces résultats et, finalement, nous présentons les ressources informatiques nécessaires à l'exécution des expériences.

7.1 Population étudiée

Nous étudions l'hypertension dans la population canadienne-française du Saguenay-Lac-Saint-Jean.

Près de 900 individus, répartis dans près de 120 familles, ont été précédemment recrutés (Hamet et al., 2005). Le principal critère d'inclusion des familles dans l'étude était la présence de deux membres hypertendus et dyslipidémiques au sein d'une fratrie. Aussi les deux parents des individus recrutés doivent être nés au Saguenay-Lac-Saint-Jean.

Le nombre de sujets recrutés varie beaucoup d'une famille à l'autre. Si on inclut les parents qui n'ont pas été recrutés mais dont la présence est nécessaire pour établir les liens généalogiques entre les sujets recrutés, la taille des familles varie de 3 à 96 individus (médiane: 7) et le nombre de générations varie de 2 à 4. En fait, l'examen d'une partie des données généalogiques du projet BALSAC de l'Université du Québec à Chicoutimi (Bouchard, 2006) nous montre que la plupart des 120 familles sont reliées les unes aux autres par au moins un de leurs membres. On pourrait donc considérer que

ces individus proviennent d'une seule grande famille de plus de 40 000 membres. Par contre, la plupart des logiciels génétiques disponibles ne permettent pas de traiter des familles de cette taille ou ne tirent pas parti de toute l'information. Certains logiciels, comme *Merlin*, sont même limités à une douzaine d'individus par famille.

7.1.1 Phénotypage

Un très grand nombre de phénotypes ont été mesurés chez les individus: pression artérielle, mesures anthropométriques (grandeur, poids, circonférence de la taille, épaisseur des plis cutanés et circonférences des extrémités), cholestérol (HDL, LDL et total) et certains marqueurs biologiques. De plus, environ 300 individus se sont soumis à un phénotypage étendu sur deux jours, dans lequel la pression artérielle a été mesurée des dizaines de fois à la suite de différents stimuli. Hamet et al. (2005) décrit le phénotypage de manière plus détaillée.

De nombreux phénotypes d'obésité ont été mesurés car « l'obésité, de façon générale, est en effet associée à une augmentation du risque d'hypertension artérielle » (Oppert, 2003).

Afin de ne pas alourdir le reste du texte, nous utilisons les abréviations et acronymes présentés au tableau 7.1. Les phénotypes sont classés par ordre alphabétique. Des références sont fournies afin de permettre au lecteur d'en connaître davantage.

Tableau 7.1: Abréviations et acronymes des phénotypes étudiés

Phénotype	***Abrév.***	***Références***
Cholestérol	CHOL	NCEP, 2002
Circonférence distale de la cuisse	CDIST	Pausova et al., 2000
Circonférence du bras	BRAS	Pausova et al., 2000
Circonférence médiale de la cuisse	CMED	Pausova et al., 2000
Circonférence proximale de la cuisse	CPROX	Pausova et al., 2000
Épaisseur du pli cutané au biceps	BICEPS	Pausova et al., 2000
Épaisseur du pli cutané au triceps	TRICEPS	Pausova et al., 2000

Phénotype	*Abrév.*	*Références*
Épaisseur du pli cutané subscapulaire	SUBSCA	Pausova et al., 2000
Épaisseur du pli cutané suprailiaque	SUPRA	Pausova et al., 2000
Glucose	GL	NCEP, 2002
Grandeur	HAUT	
Hypertension	HT	Pausova et al., 2005
Hypertension et obésité	HTOB	Pausova et al., 2005
Indice de masse corporelle	IMC	Oppert, 2003
Lipoprotéines de basse densité	LDL	NCEP, 2002
Lipoprotéines de haute densité	HDL	NCEP, 2002
Obésité	OB	Oppert, 2003
Pourcentage de gras corporel (bioimpédance)	GRASB	Pausova et al., 2000
Pourcentage de gras corporel (plis cutanés)	GRASP	Pausova et al., 2000
Pression diastolique	DBP	National Institutes of Health[8]
Pression systolique	SBP	National Institutes of Health[9]
Protéine C réactive (logarithme)	CRP	National Institutes of Health[10]
Tour de taille	TOUR	Oppert, 2003
Triglycérides	TG	NCEP, 2002

7.1.2 Génotypage

468 sujets, provenant de 76 familles, ont été génotypés par la puce Xba 50k de la compagnie Affymetrix[11]. Cette puce contient environ 58 000 SNPs répartis sur les 22 autosomes et le chromosome X. Peu de trios père-mère-enfant ont été génotypés: seuls 16 sujets génotypés ont deux parents génotypés. La plus grande famille, qui compte 96 membres, a 49 membres génotypés. Par ailleurs, la totalité des 300 sujets pour lesquels un phénotypage complet a été effectué ont été génotypés.

8 http://www.nlm.nih.gov/medlineplus/ency/article/003398.htm
9 http://www.nlm.nih.gov/medlineplus/ency/article/003398.htm
10 http://www.nlm.nih.gov/medlineplus/ency/article/003356.htm
11 http://www.affymetrix.com/products/arrays/specific/100k.affx

7.2 Expériences

Nous avons conçu huit expériences afin de comparer notre méthode de correction à la correction de Bonferroni dans plusieurs contextes. Nous avons fait varier les paramètres suivants:

1. Le nombre de SNPs étudiés simultanément: entre 5 et 57 000.
2. Le nombre de phénotypes étudiés simultanément: entre 1 et 16.
3. La distribution des phénotypes: binomiale ou normale.
4. La présence ou non de liaison entre les SNPs et les phénotypes.

Tableau 7.2: Paramétrisation des expériences.

Section	***Titre***	***H_0***	***Modèles***	***N***
7.2.1	Phénotypes héritables	H_{02}	*m*: *a*	1 000
7.2.2	Hypertension avec obésité	H_{02}	*m*: *a*	1 000
7.2.3	Phénotypes anthropométriques	H_{02}	*m*: *a*	1 000
7.2.4	Protéine C réactive	H_{01} et H_{02}	*m*, *b*: *a*, *d*, *g*, *r*	1 000
7.2.5	Gènes candidats de CRP	H_{02}	*m*, *b*: *a*, *d*, *g*, *r*	100 000
7.2.6	Gène FATP6 et syndrome métabolique	H_{01} et H_{02}	*m*: *a*	10 000
7.2.7	Cardiopathie coronarienne	H_{01} et H_{02}	*m*, *b*: *a*, *d*, *r*	10 000
7.2.8	Gras corporel par plis cutanés	H_{01}	*m*: *a*	100 000

La colonne H_0 contient l'hypothèse nulle testée: « pas d'association ni de liaison » (H_{01}) ou « pas d'association en présence de liaison » (H_{02}) (section 3.5). H_{01} est testée par la commande *fbat* et H_{02} par *fbat -e*. La colonne Modèles indique le ou les modèles génétiques testés (section 3.3): additif, dominant, récessif ou génotypique, représentés respectivement par les lettres *a*, *d*, *r* et g, et le ou les modes alléliques: un test par allèle (bi-allélique, noté *b*) ou un test par SNP (multi-allélique, noté *m*). La colonne N représente le nombre de simulations: l'augmenter rétrécit l'intervalle de confiance sur la valeur p corrigée (section 5.1.1) au prix d'un temps de calcul proportionnellement plus long.

Le tableau 7.2 résume les huit expériences. On y retrouve les paramètres fournis à *FBAT* pour chaque expérience: l'hypothèse nulle, les modèles et modes génétiques et le nombre de simulations. Les autres variables d'intérêt qui n'apparaissent pas dans ce tableau sont les phénotypes et covariables utilisés, les SNPs testés et les groupes choisis. Dans toutes les expériences, le seuil de signficativité après correction pour tests multiples a été fixé à 0,05 et le nombre minimum de familles informatives est de 5.

7.2.1 Phénotypes héritables

Cette expérience teste l'association entre 16 phénotypes anthropométriques (BICEPS, BRAS, CDIST, CMED, CPROX, GRASB, GRASP, HAUT, IMC, SUPRA, TOUR, TRICEPS) et métaboliques (CHOL, HDL, LDL, TG) et les 22 autosomes, soit environ 57 000 SNPs. Les phénotypes ont été choisis en fonction des résultats de liaison très prometteurs rapportés par Hamet et al. (2005, figure 5) et du fait qu'ils sont très héritables (Hamet et al., 2005, tableau 1).

Mille simulations ont été effectuées. Bien que ce nombre puisse paraître faible, il est tout de même assez élevé pour que l'intervalle de confiance à 95% d'une valeur p corrigée de 0,03 soit [0,019; 0,041]; en d'autres termes, l'intervalle de confiance n'inclut pas le seuil de 0,05. Si on obtenait une telle valeur p corrigée, on pourrait conclure qu'elle est inférieure à 0,05 et déclarer l'association statistiquement significative.

Notons que cette correction pour tests multiples risque d'être très sévère pour cette expérience, en raison du grand nombre de SNPs (57 000) et de phénotypes (16) testés. Il est possible de diminuer le nombre de SNPs testés et, possiblement, obtenir de meilleures valeurs p corrigées. On peut choisir les SNPs les plus prometteurs selon la statistique de test PBAT (Van Steen et al., 2005), se limiter à une région liée au phénotype (Roeder et al., 2006) ou se limter à des gènes candidats (Barroso et al., 2003). Dans les trois cas, il s'agit d'utiliser une information statistiquement indépendante du test d'association afin de se limiter aux SNPs ayant une probabilité d'association à priori élevée. Les expériences suivantes se basent toutes sur cette idée.

7.2.2 Hypertension avec obésité

Pausova et al. (2005) ont étudié l'hypertension chez des familles obèses et non-obèses du Saguenay-Lac-Saint-Jean. Les familles étudiées sont les mêmes que celles décrites au début de ce chapitre. Une famille est déclarée obèse si au moins 70% de ses membres ont un indice de masse corporelle (IMC) d'au moins 27 kg/m^2. L'analyse de liaison chez toutes les familles a permis de détecter des liaisons de significativité modeste sur le chromosome 1 près de 40 cM pour HT et HTOB (LOD scores respectifs de 1,9 et 2,3). Or, l'étude séparée des familles obèses a permis de détecter une liaison de significativité beaucoup plus forte au même endroit pour HT, OB et HTOB (LOD scores respectifs de 2,9, 3,3 et 3,5). De plus, selon Pausova et al. (2005), au moins trois études indépendantes ont rapporté des liaisons significatives ou suggestives entre des phénotypes d'obésité ou d'hypertension et cette région. Finalement, cette région contient plusieurs gènes candidats. Ces résultats nous amènent à penser qu'une analyse d'association sur les phénotypes HT, OB et HTOB et les SNPs de cette région pourrait rapporter des associations significatives.

Les phénotypes HT et HTOB sont binaires. Le phénotype HT est VRAI si le sujet a reçu un diagnostique médical d'hypertension, FAUX sinon. Le phénotype HTOB est VRAI si HT est vrai et si l'IMC du sujet est supérieur ou égal à 27, FAUX sinon. Lorsque nous n'avons pas toute l'information phénotypique pour un sujet, nous lui assignons une valeur manquante.

La région du chromosome 1 comprise entre 20 et 60 cM, soit 20 cM en aval et 20 cM en amont du pic rapporté par Pausova et al. (2005), est étudiée. Elle comprend 191 SNPs.

7.2.3 Phénotypes anthropométriques

Hamet et al. (2005) ont rapporté plusieurs liaisons statistiquement significatives entre un grand nombre de phénotypes anthropométriques et métaboliques et la région du chromosome 1 comprise entre 160 et 240 cM. L'intérêt de cette région réside dans la significativité des liaisons observées (un des LOD scores atteint 4, alors que Lander et

Kruglyak (1994) recommandent d'utiliser un seuil de 3,3 pour déclarer une liaison statistiquement significative) et dans le grand nombre de signaux suggestifs ou significatifs au même endroit.

L'expérience porte sur les cinq phénotypes ayant fourni les meilleurs LOD scores dans la région (CDIST, CPROX, GRASP, SUPRA et TRICEPS) et les 1637 SNPs de la région.

7.2.4 Protéine C réactive

La protéine C réactive (*C-reactive protein*, CRP) est un marqueur d'inflammation produit par le foie. Ridker et al. (2000) ont constaté que le taux sanguin de la CRP prédit en partie le risque de maladies cardiovasculaires chez les femmes post-ménopausées, même chez celles dont le niveau de LDL est jugé adéquat. Dans une étude portant sur plus de 6000 hommes et femmes, Danesh et al. (2004) ont trouvé qu'un niveau plus élevé de CRP est associé à un risque accru de maladies cardiovasculaires. Ils ont démontré via une méta-analyse de 22 études que la CRP est un prédicteur modéré de maladies cardiovasculaires. Selon Tremblay (2007), la CRP pourrait également être un prédicteur précoce de l'hypertension.

Une analyse de liaison dans deux populations indépendantes, soit un échantillon de 500 familles d'Europe de l'Ouest et notre échantillon de familles canadiennes-françaises, a démontré l'existence d'une liaison entre les marqueurs microsatellites se situant entre 125 et 140 cM du chromosome 10 et la CRP (Broeckel et al., 2007). Dans les familles d'Europe de l'Ouest, l'analyse de liaison rapporte un LOD score de 3,15.

En conséquence, une expérience a été mise au point afin de déterminer si un des 680 SNPs de la région 120-160 cM du chromosome 10 est associé à la CRP. Seules les valeurs de CRP inférieures ou égales à 5 ont été incluses car les valeurs élevées de CRP reflètent une inflammation aiguë plutôt que chronique (Tremblay, 2007). Puisqu'une liaison a été détectée entre la CRP et cette région, nous avons testé l'hypothèse nulle H_{02} (« pas d'association en présence de liaison »).

7.2.5 Gènes candidats de CRP

L'étude de Broeckel et al. (2007) a permis de générer une liste de gènes candidats de la CRP, en plus de recenser les gènes déjà supposés moduler la CRP. En plus de la liaison sur le chromosome 10, des liaisons moins significatives ont été détectées sur les chromosomes 2 (LOD score de 2,5 chez les hommes) et 5 (LOD score de 2,2). Les principaux gènes de ces régions de liaison sont les suivants:

- Chromosome 2: TGFA, IL1A, IL1B, IL1R1, IL1R2 et IL18R1
- Chromosome 5: TGFB1, IL3, IL4, IL5, IL9, IL13, CSF2 et CD14
- Chromosome 10: ADAM8, PRLHR et ADRB1

Broeckel et al. (2007) ont également recensé dans la littérature les principaux gènes candidats de CRP: ce sont les gènes CRP, IL6, TLR4 et IL1.

Nous avons cherché les SNPs qui se trouvent à l'intérieur des 21 gènes et qui sont présents sur la puce Xba. Nous avons identifié 26 SNPs dans 7 gènes: TGFA (10 SNPs), IL1R1 (1), IL1R2 (3), IL18R1 (3), IL9 (1), PRLHR (6) et IL6 (2). Le faible nombre de SNPs testés a permis d'exécuter 100 000 simulations.

7.2.6 Gène FATP6 et syndrome métabolique

Plusieurs études impliquent le gène FATP6 (aussi connu sous le nom SLC27A6) dans les maladies cardiovasculaires reliées aux lipides (Gimeno et al., 2003). Ce gène, situé sur le chromosome 5, code pour une protéine de 620 acides aminés.

Une expérience a été conduite afin de déterminer si un des 5 SNPs de la puce Xba qui se trouvent dans le gène est associé à un des phénotypes impliqués dans le syndrome métabolique[12]: DBP, GL, HDL, SBP et TG.

12 http://www.theses.ulaval.ca/2004/22151/22151.html

7.2.7 Cardiopathie coronarienne

Une étude récente portant sur un échantillon indépendant de la population canadienne-française du Saguenay-Lac-Saint-Jean a démontré une liaison et des associations statistiquement significatives entre des gènes candidats de la cardiopathie coronarienne et des phénotypes intermédiaires (Paré et al., 2007). En particulier, la région 1p36.22 (de 8,9 Mb à 11,8 Mb sur le chromosome 1) est liée au HDL (LOD score de 2,52, valeur p empirique corrigée pour tous les tests de liaison < 0,05).

Ces résultats sont particulièrement intéressants car ils ont été obtenus chez la population du Saguenay-Lac-Saint-Jean: souvent, des différences entre populations font en sorte que des associations trouvées dans une population ne sont pas répliquées dans une autre population (voir par exemple Serena et al., 2004). Nous avons donc étudié l'association entre les 14 SNPs de la région de liaison 1p36.22 et le HDL.

7.2.8 Gras corporel par plis cutanés

En réexaminant la région de liaison du chromosome 1 identifiée par Hamet et al. (2005), il ressort que les cinq phénotypes de pli cutané (BICEPS, GRASP, SUBSCA, SUPRA, et TRICEPS) présentent des pics de liaison avec marqueurs microsatellites qui varient en terme de significativité statistique mais qui sont tous situés pratiquement au même endroit (fgure 7.1). De plus l'analyse de liaison effectuée sur les SNPs révèle que, parmi les 5 phénotypes de pli cutané, seul GRASP présente un LOD score de plus de 3,0 au même endroit que le pic des microsatellites (figure 7.2).

Une expérience a été effectuée afin de déterminer si un SNP près du pic est associé à GRASP. Les 59 SNPs distants de moins de 1 cM du pic de liaison des SNPs ont été retenus. Le choix de 1 cM est arbitraire mais se veut un compromis entre, d'une part, les chances d'inclure le SNP associé au phénotype et, d'autre part, le fait que plus le nombre de SNPs inclus augmente, plus la correction est sévère. 100 000 simulations ont été effectuées.

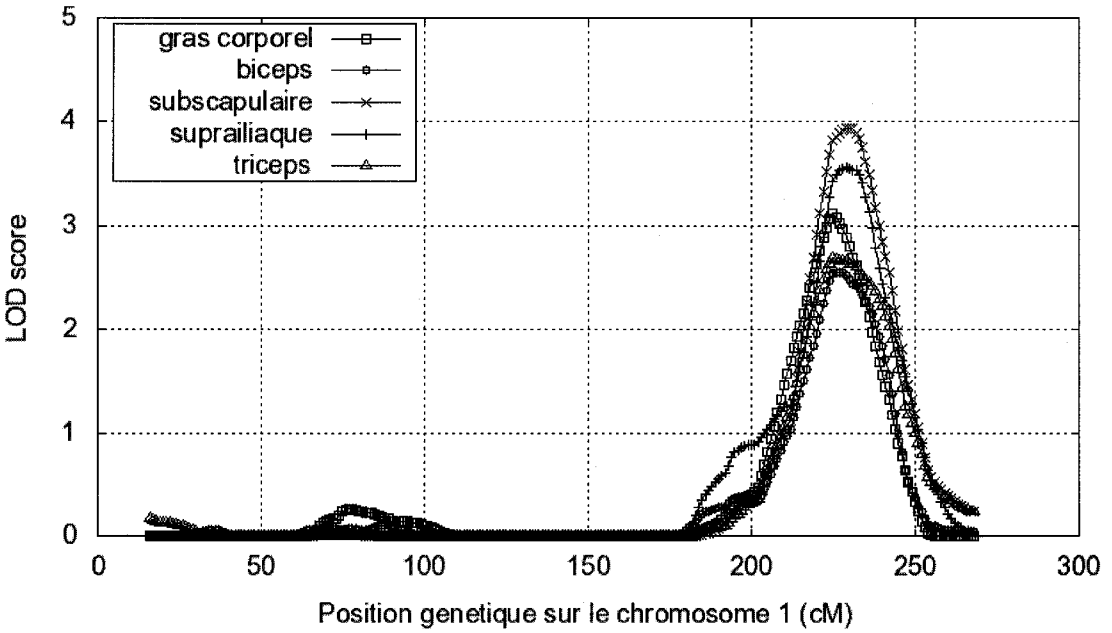

Figure 7.1: Liaison entre les plis cutanés et les microsatellites

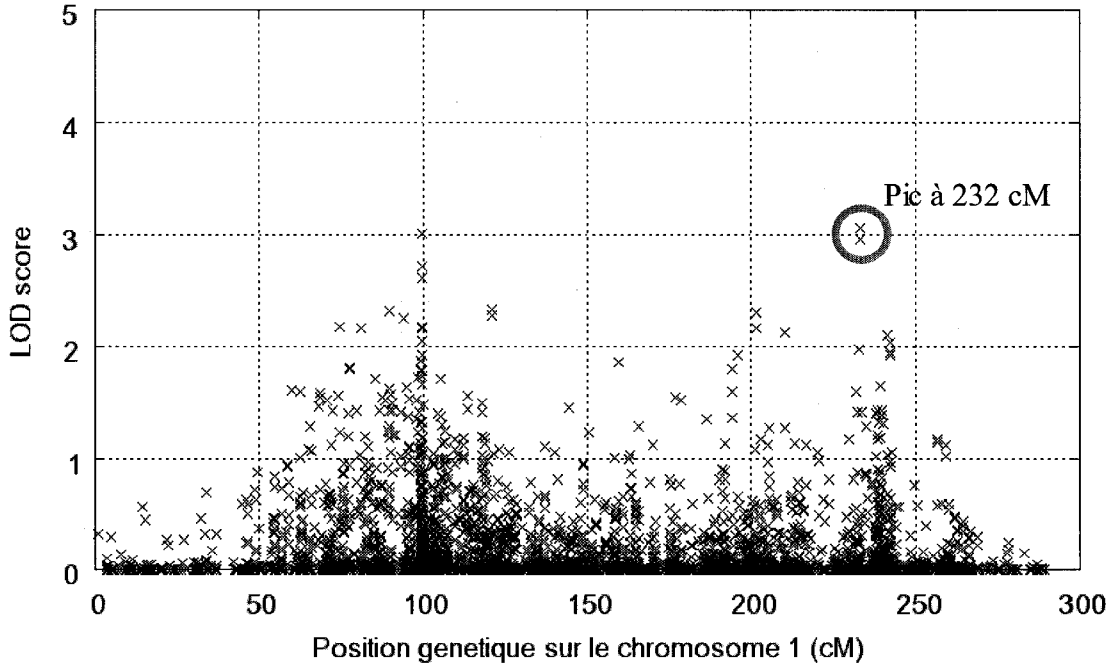

Figure 7.2: Liaison entre GRASP et les SNPs

7.3 Résultats

Les résultats des expériences sont rapportés au tableau 7.3. Il y a un bloc par expérience. À l'intérieur d'un bloc, chaque ligne représente un phénotype. Les phénotypes sont classés en fonction de leur meilleure valeur p observée, du meilleur résultat au pire. La colonne p_{obs}, représente la valeur p de la meilleure association pour ce phénotype, sim_1 et $bonf_1$ représentent la valeur p corrigée de la meilleure association lorsqu'on considère chaque phénotype séparément (selon notre méthode et selon la correction de Bonferroni, respectivement) alors que sim_n et $bonf_n$ représentent la valeur p corrigée de la meilleure association lorsqu'on considère tous les phénotypes simultanément. Formellement, sim_1 est la probabilité d'observer un résultat FBAT aussi extrême lorsqu'aucun SNP n'est associé au phénotype considéré, alors que sim_n est la probabilité d'observer un résultat FBAT aussi extrême lorsqu'aucun SNP n'est associé à aucun des phénotypes étudiés. Nous présentons sim_1 et sim_n parce que le choix de la *famille* de tests est subjectif (Westfall & Young, 1993, p. 21).

Tableau 7.3: Résultats d'association

Section	***Phénotype***	p_{obs}	sim_1	sim_n	$bonf_1$	$bonf_n$
7.2.1	SUPRA	$9{,}1\times10^{-5}$	0,03	0,45	1	1
	CDIST	$1{,}2\times10^{-4}$	0,06	0,64	1	1
	IMC	$1{,}9\times10^{-4}$	0,19	0,91	1	1
	CHOL	$1{,}9\times10^{-4}$	0,19	0,91	1	1
	GRASP	$2{,}0\times10^{-4}$	0,18	0,94	1	1
	BRAS	$3{,}0\times10^{-4}$	0,40	1	1	1
	CMED	$3{,}6\times10^{-4}$	0,44	1	1	1
	TRICEPS	$4{,}2\times10^{-4}$	0,59	1	1	1
	LDL	$4{,}4\times10^{-4}$	0,70	1	1	1
	CPROX	$4{,}6\times10^{-4}$	0,53	1	1	1
	TOUR	$5{,}3\times10^{-4}$	0,79	1	1	1
	TG	$6{,}4\times10^{-4}$	0,68	1	1	1

Section	*Phénotype*	p_{obs}	sim_1	sim_n	$bonf_1$	$bonf_n$
	HDL	$7{,}4\times10^{-4}$	0,95	1	1	1
	BICEPS	$7{,}4\times10^{-4}$	0,83	1	1	1
	HAUT	$8{,}1\times10^{-4}$	0,94	1	1	1
	GRASB	$1{,}0\times10^{-3}$	0,96	1	1	1
7.2.2	OB	0,002	0,04	0,12	0,29	0,87
	HTOB	0,003	0,10	0,26	0,44	1
	HT	0,007	0,40	0,76	1	1
7.2.3	SUPRA	0,002	0,37	0,98	1	1
	CDIST	0,003	0,57	1	1	1
	TRICEPS	0,003	0,55	1	1	1
	CPROX	0,003	0,51	1	1	1
	GRASP	0,004	0,66	1	1	1
7.2.4 (H_{01})	CRP	0,002	0,17	0,17	1	1
7.2.4 (H_{02})	CRP	0,0002	0,10	0,10	0,14	0,14
7.2.5	CRP	0,005	0,07	0,07	0,13	0,13
7.2.6	GL	0,02	0,05	0,25	0,10	0,50
	HDL	0,07	0,24	0,70	0,35	1
	TG	0,39	0,86	1	1	1
	SBP	0,39	0,86	1	1	1
	DBP	0,62	0,98	1	1	1
7.2.7	HDL	0,04	0,45	0,45	0,59	0,59
7.2.8	GRASP	0,0037	0,037	0,037	0,22	0,22

Ces résultats sont intéressants à plus d'un égard. D'abord, nous avons trouvé une association statistiquement significative après correction pour tests multiples (expérience 7.2.8). Ensuite, les corrections par simulation sont en général beaucoup moins sévères que les corrections par Bonferroni. Aussi, le test FBAT appliqué à nos données manque de puissance. Finalement, sept des huit expériences n'ont pas fourni de résultats statistiquement significatifs après correction pour tests multiples. Reprenons

ces découvertes une à une.

7.3.1 Association statistiquement significative

Pour l'expérience 7.2.8, le SNP le plus fortement associé est rs10494966 (p_{obs} = 0,0037), situé sur le chromosome 1 à 212 068 043 pb selon la version 127 de la base de données dbSNP (identificateur Affymetrix 1648544). Seules 3710 des 100 000 simulations ont produit une association de meilleure valeur p que celle observée. La valeur p corrigée est donc de 0,037. L'erreur standard est de 0,0006. L'intervalle de confiance à 95% est [0,036; 0,038]. *L'association entre le SNP rs10494966 et le pourcentage de gras corporel calculé à partir des plis cutanés est significative après correction pour tests multiples.* En comparaison, la valeur p corrigée par la méthode de Bonferroni est 0,22.

Knapp et Becker (2004) ont démontré que des erreurs de génotypage mêmes rares pouvaient avoir un grand impact dans les tests d'association génomique familiale basés sur les haplotypes (test HS-TDT). Pour un taux d'erreur aussi faible que 1%, jusqu'à 57% des hypothèses nulles vraies sont rejetées pour un taux nominal α de 5%. Le test HS-TDT peut donc être extrêmement libéral en cas d'erreurs de génotypage. Le test TDT original (implanté dans *FBAT*) souffre du même problème. Dans le contexte d'études cas-témoins, Moskvina et al. (2006) ont démontré que des taux d'erreurs de génotypage différents chez les cas et les témoins pouvaient augmenter la probabilité de rejeter une hypothèse nulle vraie. Quatre vérifications ont donc été menées afin de s'assurer que l'association entre rs10494966 et GRASP n'est pas due à des erreurs.

Premièrement, un faible taux de génotypage à ce SNP pourrait indiquer un nombre élevé d'erreurs de génotypage ou encore un génotypage sélectif (Affymetrix, 2006). Ce taux représente la proportion des individus génotypés qui ont été « appelés » par l'algorithme de génotypage pour ce SNP (pour les autres individus, l'algorithme ne peut déterminer leur génotype à ce SNP). Or, ce taux est 99,6%, ce qui est très élevé lorsqu'on le compare au seuil minimum de 95% généralement utilisé dans les études d'association (Diabetes Genetics Initiative et al., 2007; Sladek et al., 2007).

Deuxièmement, un SNP qui n'est pas en équilibre de Hardy-Weinberg peut également indiquer des erreurs de génotypage ou encore que le SNP se situe dans une zone de variabilité du nombre de copies (*copy number variation*, CNV; Redon et al., 2006). Dans le cas particulier des puces d'Affymetrix, l'algorithme de génotypage a de la difficulté à identifier correctement les génotypes hétérozygotes dans les CNV: dans ce cas, l'algorithme rapporte des valeurs manquantes ou des génotypes homozygotes (Affymetrix, 2006). Un test exact des proportions d'Hardy-Weinberg (hypothèse nulle: le marqueur est en équilibre d'Hardy-Weinberg) rapporte une valeur p de 0,79: il n'y a donc aucune raison de penser que le SNP n'est pas en équilibre de Hardy-Weinberg. Les seuils minima généralement utilisés varient de 10^{-6} à 10^{-3}.

Troisièmement, une fréquence de l'allèle mineur très faible pourrait fournir des valeurs p très bonnes uniquement parce que l'approximation asymptotique utilisée par *FBAT* pour calculer la valeur p ne serait pas valide. Notons qu'on se protège partiellement de cette possibilité d'erreur en imposant dans tous nos tests un nombre minimal de 5 familles informatives, faute de quoi le test n'est pas effectué. La fréquence de l'allèle mineur est de 9,9%, ce qui est supérieur au seuil minimum de 1% généralement utilisé pour exclure un SNP d'un test d'association.

Finalement, la commande *allfreq* appelée à partir du logiciel *SOLAR* tente de détecter les erreurs de génotypage en s'assurant que les génotypes des membres d'une famille respectent les lois de Mendel, c'est-à-dire que chaque individu reçoit un allèle de son père et un allèle de sa mère. Cette commande peut tirer profit de l'information contenue dans une famille étendue et ainsi détecter davantage d'erreurs que si elle se concentrait sur chaque famille nucléaire séparément. Or, la commande *allfreq* ne rapporte aucune erreur de génotypage pour rs10494966.

En résumé, les quatre vérifications menées indiquent que l'association entre le SNP rs10494966 et GRASP n'est pas due à une erreur de génotypage. À l'instar de Lazzeroni et Lange (1998, p. 69), *nous concluons que le SNP rs10494966 est soit directement responsable de la valeur du pourcentage de gras corporel chez les individus de la*

population étudiée, soit en LD avec un gène qui module le pourcentage de gras corporel.

Le SNP rs10494966 se trouve à l'intérieur d'un intron du gène putatif Hs.368496. Le gène connu le plus proche est PROX1, situé entre 212 228 483 et 212 276 389 pb, soit à 160 kb du SNP. Plusieurs études impliquent PROX1 dans le cancer[13].

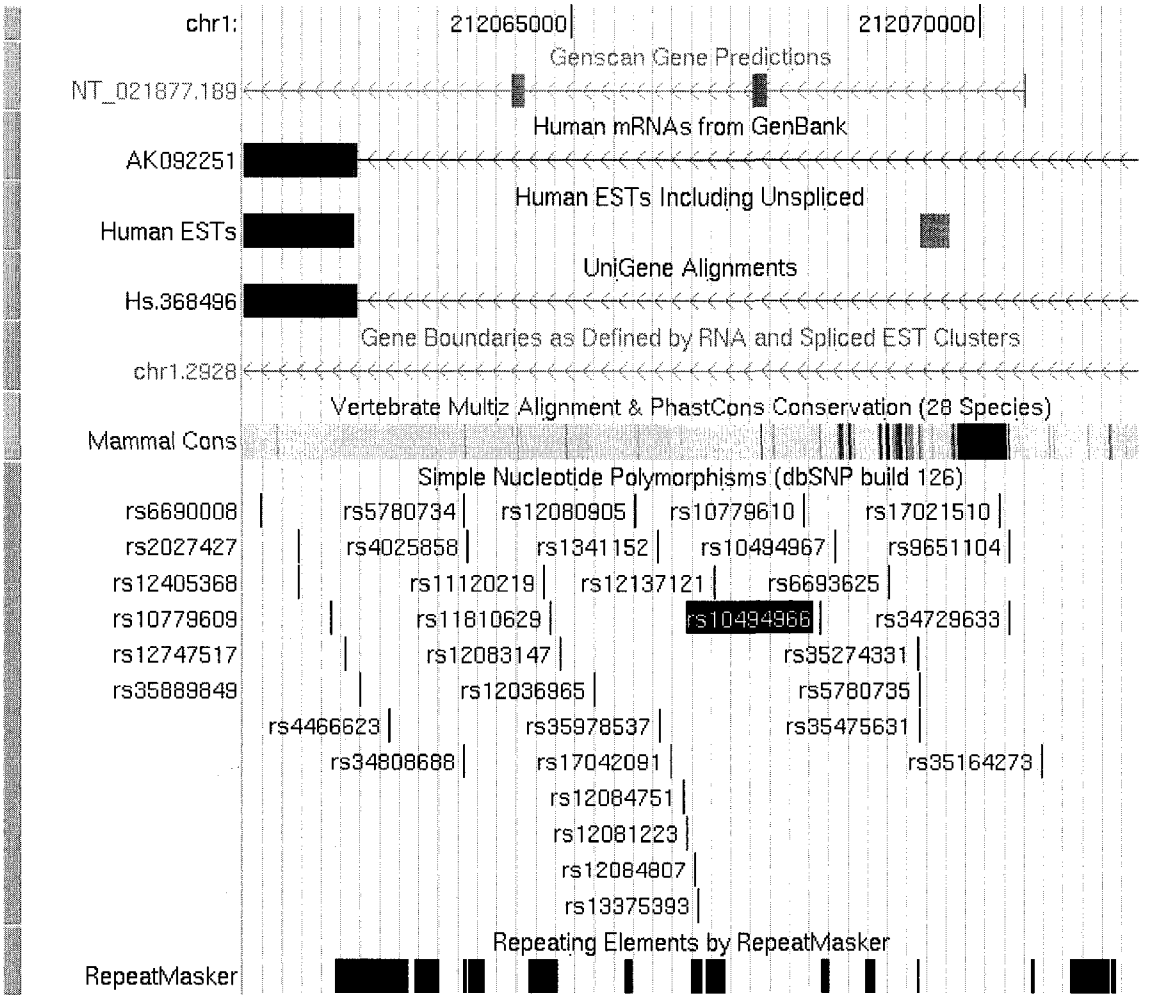

Figure 7.3: Contexte génomique de rs10494966

Les éléments génomiques d'intérêt du chromosome 1 entre les positions 212 061 000 et 212 071 939 paires de bases. Le diagramme a été produit par le fureteur génomique de l'Université de Californie à Santa Cruz (http://genome.ucsc.edu). La 5ème rangée représente le gène putatif Hs.368496 et montre que le SNP rs10494966 se trouve dans l'intron du gène putatif, i.e. dans une région non codante.

13 Ces informations proviennent de la version 36.2 du génome humain de NCBI.

Le fait que le SNP ne se retrouve pas dans un gène connu n'invalide pas l'association trouvée. En effet, il est possible qu'un polymorphisme se situant à l'extérieur d'une région codante de l'ADN ait un effet physiologique. Les régions non codantes représentent près de 99% du génome des humains (International Human Genome Sequencing Consortium, 2004). Selon des estimations conservatrices, au moins 60% des régions non codantes sont transcrites en ARN. Il est de plus en plus évident que cet ARN non codant joue un rôle biologique important. Par exemple, les ARN non-codants Xist et Tsix contrôlent l'inactivation du chromosome X chez les mammifères. D'autres ARN non-codants ont été associés aux syndromes de Prader-Willi et d'Angelman (Mattick & Makunin, 2006). D'ailleurs, des études récentes ont mis en évidence des associations significatives entre des maladies et des SNPs se situant dans des régions non codantes: une des associations rapportée par Diabetes Genetics Initiative et al. (2007) se trouve dans un bloc de LD pour lequel on ne connaît aucun gène, alors qu'une des associations de Maller et al. (2006) se trouve dans une région non codante d'un gène, ce qui les amène à conclure que « des allèles ayant un effet substantiel sur une maladie répandue peuvent être trouvés à l'extérieur des gènes candidats et à l'extérieur des régions codantes. »

7.3.2 Simulations plus puissantes que Bonferroni

Dans toutes les expériences, les corrections par simulations ne sont pas plus sévères que les corrections de Bonferroni; dans la plupart des expériences, elles sont beaucoup moins sévères. Par exemple, pour l'expérience 7.2.1, la valeur p corrigée par Bonferroni correspondant à une valeur p observée de $9{,}1\times10^{-5}$ est de 1 lorsqu'on considère 57 000 SNPs et un seul phénotype. En comparaison, notre meilleure valeur p corrigée est de 0,03: dans ce cas, la correction de Bonferroni est 33 fois plus sévère que notre correction.

Notre méthode de correction est moins sévère indépendamment du nombre de SNPs et du nombre de phénotypes étudiés simultanément (tableau 7.4), du type de phénotype

étudié (l'expérience 7.2.2 examinait deux phénotypes binaires) et de la présence ou non de liaison entre les SNPs et les phénotypes. Lorsqu'on considère chaque phénotype séparément, les corrections par simulations sont entre 1,3 et 33 fois moins sévères. Lorsqu'on considère les phénotypes collectivement, les corrections sont jusqu'à 7 fois moins sévères.

Tableau 7.4: Comparaison des simulations à Bonferroni

Section	***SNPs***	***Phénos***	$ratio_1$	$ratio_n$
7.2.1	57 000	16	33,3	2,2
7.2.2	191	3	7,3	7,3
7.2.3	1 637	5	2,7	1,0
7.2.4 (H_{01})	680	1	5,9	5,9
7.2.4 (H_{02})	680	1	1,4	1,4
7.2.5	26	1	1,9	1,9
7.2.6	5	5	2,0	2,0
7.2.7	14	1	1,3	1,3
7.2.8	59	1	6,0	6,0

La colonne SNPs indique le nombre de SNPs testés dans l'expérience, Phénos indique le nombre de phénotypes testés, $ratio_1$ est le rapport entre $bonf_1$ et sim_1 (voir le tableau 7.3) et $ratio_n$ est le rapport entre $bonf_n$ et sim_n.

7.3.3 Faible puissance

Pour l'expérience 7.2.1, les meilleures valeurs p observées varient de $9{,}1\times10^{-5}$ (SUPRA) à $1{,}0\times10^{-3}$ (GRASB). 867 074 tests FBAT ont été effectués.

Les meilleures valeurs p observées semblent anormalement élevées. En effet, même en l'absence d'association, une proportion d'environ $9{,}1\times10^{-5}$ des 867 074 tests, soit 79 tests, devraient rapporter une valeur p observée inférieure ou égale à $9{,}1\times10^{-5}$, en supposant les tests indépendants. S'il y avait des associations, cette proportion devrait

être encore plus élevée. Le fait qu'une seule valeur p observée soit inférieure ou égale à $9{,}1\times10^{-5}$ suggère que le test FBAT appliqué à nos données n'est pas puissant.

La figure 7.4 présente les 867 074 valeurs p observées en fonction des valeurs p attendues en absence d'association. La meilleure valeur p observée, $9{,}1\times10^{-5}$, se situe à la gauche du diagramme. La ligne diagonale représente la distribution attendue des valeurs p. Les échelles du diagramme sont logarithmiques afin de mettre l'emphase sur les petites valeurs p. Le fait que la courbe observée se situe au-dessus de la courbe théorique indique que le test FBAT appliqué à nos données est nettement conservateur.

Ces résultats conservateurs pourraient s'expliquer par le fait que l'expérience 7.2.1 a testé l'hypothèse H_{02} (« pas d'association en présence de liaison »), qui est plus conservatrice que l'hypothèse H_{01} (« pas d'association ni de liaison »).

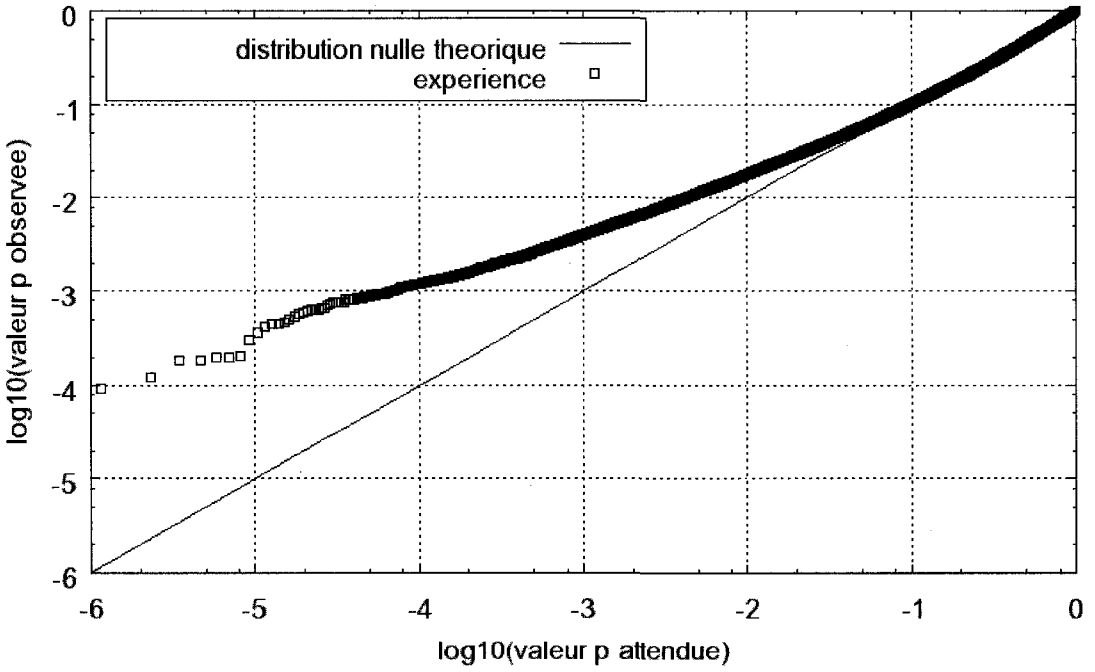

Figure 7.4: Distribution des 867 074 valeurs p observées de l'expérience 7.2.1

7.3.4 Nécessité de guider la recherche

Pour l'expérience 7.2.1, lorsqu'on considère les phénotypes individuellement, la meilleure association pour SUPRA ($sim_l = 0,03$) est significative à un seuil de 0,05. Par contre, lorsqu'on considère les phénotypes collectivement (sim_n), aucune association n'est significative pour un seuil de 0,05: la meilleure association de SUPRA correspond à une valeur p ajustée de 0,45. Pour 11 des 16 phénotypes, la meilleure association correspond à une valeur p ajustée de 1: lorsqu'aucune association n'existe entre les phénotypes et les génotypes étudiés, au moins un test FBAT rapportera une valeur p meilleure que toutes celles observées pour ces phénotypes.

Ces résultats sont décevants, surtout lorsqu'on considère que les phénotypes étudiés ont déjà démontré des signaux de liaison très forts en plusieurs endroits, notamment sur le chromosome 1, et qu'ils sont très héritables. En particulier, HAUT est un phénotype extrêmement héritable (corrélation familiale calculée par FCOR comprise entre 0,56 et 0,68 chez les hommes et entre 0,46 et 0,47 chez les femmes; Hamet et al., 2005). Or, la meilleure association pour HAUT rapporte une valeur p observée de $8,1\times10^{-4}$, ce qui n'est pas significatif ($sim_l = 0,94$, $sim_m = 1$).

Le même phénomène se produit à l'expérience 7.2.2: lorsqu'on corrige uniquement pour le nombre de SNPs, la meilleure valeur p corrigée est 0,04 (donc significative), alors que lorsqu'on corrige également pour le nombre de phénotypes, l'association n'est pas significative ($sim_n = 0,12$). En d'autres termes, à la question « quelle est la probabilité d'obtenir une statistique FBAT au moins aussi bonne que la meilleure statistique FBAT observée lorsqu'il n'existe aucune association entre les 191 SNPs et les trois phénotypes? », la réponse est 0,12. Aucune hypothèse nulle ne peut donc être rejetée à un seuil de significativité de 0,05.

En tout, sur 8 expériences, 4 produisent au moins une association significative lorsqu'on considère chaque phénotype individuellement, mais seulement une est significative lorsque tous les tests sont pris en compte simultanément.

Il est à noter que des valeurs p observées plus grandes ne résultent pas nécessairement en des valeurs p corrigées plus grandes. Par exemple, dans l'expérience 7.2.1, la meilleure valeur p observée pour le LDL ($4{,}4\times10^{-4}$) est plus petite que celle pour CPROX ($4{,}6\times10^{-4}$) mais les valeurs p ajustées sont respectivement 0,70 et 0,53. Cela signifie qu'il est plus rare d'obtenir une valeur p de $4{,}6\times10^{-4}$ pour CPROX que d'obtenir une valeur p de $4{,}4\times10^{-4}$ pour le LDL, lorsqu'aucun SNP n'est associé aux phénotypes. Une distribution différente des valeurs phénotypiques et un nombre différent de sujets phénotypés sont deux explications possibles de ce phénomène. En effet, dans beaucoup de tests statistiques, des données asymétriques donnent de meilleures valeurs p que des données symétriques. Aussi, lorsque le nombre d'observations est faible, la puissance des tests statistiques est faible (Ostle et al., 1996).

Deux hypothèses nulles ont été testées durant l'expérience 7.2.4. Les meilleures valeurs p observées sont de 0,0002 pour H_{01} (« pas d'association ni de liaison ») et 0,002 pour H_{02} (« pas d'association en présence de liaison »), soit un ordre de grandeur de différence. Or, les valeurs p corrigées se ressemblent: 0,17 et 0,10, respectivement. Cela indique que, pour cette expérience, il est plus facile d'obtenir une valeur p de 0,0002 pour H_{01} qu'une valeur p de 0,002 pour H_{02}, en absence d'association. Cette constatation est importante car, habituellement, une valeur p de 0,0002 est interprétée comme étant une preuve plus forte contre l'hypothèse nulle qu'une valeur p de 0,002.

Dans l'expérience 7.2.3, on constate que les deux phénotypes pour lesquels les valeurs p observées sont les meilleures, soit SUPAR et CDIST, sont également les deux meilleurs phénotypes lorsque tout le génome est analysé (expérience 7.2.1). Les meilleures valeurs p observées sont moins bonnes lorsque seule une région du chromosome 1 est analysée; par contre, moins de marqueurs sont analysés et la correction est moins sévère. Malgré tout, les associations ne sont pas significatives.

Les résultats que nous avons obtenus montrent qu'il est bénéfique de guider la recherche afin de réduire la sévérité de la correction. En effet, ce n'est que lorsque nous nous sommes limités à une courte région autour d'un pic de liaison que nous avons obtenu

une association statistiquement significative après correction pour tests multiples. Pour que les valeurs p corrigées soient valides, les informations utilisées pour guider la recherche et les tests d'association doivent être statistiquement indépendants les uns des autres. Nous nous sommes servis des résultats d'analyses de liaison, effectuées sur notre échantillon ou sur d'autres échantillons, et des résultats d'analyses d'association provenant d'autres études.

7.4 Temps de calcul et mémoire

Nous concluons ce chapitre par la présentation des temps de calcul et l'utilisation de la mémoire vive.

Le temps de calcul est proportionnel au nombre de simulations effectuées et varie de façon cubique par rapport au nombre de SNPs analysés.

Les régions qui contiennent peu de SNPs ont pu être analysées rapidement. Par exemple, l'expérience 7.2.2, qui étudiait 191 SNPs et pour laquelle 1000 simulations ont été effectuées, n'a nécessité que 8 minutes grâce à l'utilisation de 16 processeurs. L'expérience 7.2.8 (59 SNPs et 100 000 simulations) n'a nécessité qu'un peu plus de 3 heures sur 17 processeurs.

En revanche, l'expérience 7.2.1 (57 000 SNPs et 1000 simulations) a nécessité 57 heures de calcul sur 16 processeurs, soit 912 heures mono-processeur. Il est toutefois bon de remarquer que même un nombre de simulations qui peut paraître faible (par exemple 1000 simulations) peut être adéquat: si l'intervalle de confiance à 95% de la valeur p corrigée ne contient pas le seuil nominal fixé, il est inutile d'exécuter davantage de simulations. Par exemple, l'intervalle de confiance à 95% d'une valeur p corrigée à partir de 1000 simulations de 0,03 est [0,019; 0,041].

L'étape la plus gourmande en mémoire et en temps de calcul est la génération des haplotypes des fondateurs. Nous avons divisé les chromosomes en grands blocs afin de pouvoir utiliser l'architecture existante (soit des serveurs équipés de 1,5 Go de mémoire vive). Les autres étapes ne demandent que quelques dizaines de Mo chacune.

Chapitre 8: Discussion

Dans ce chapitre nous rappelons les expériences effectuées, comparons la méthode proposée à la correction de Bonferroni, soulignons l'association statistiquement significative que nous avons découverte grâce à notre méthode, résumons les principales propriétés de notre méthode, suggérons une méthode de rééchantillonnage alternative et discutons des menaces à la validité interne et externe.

8.1 Rappel des expériences

Huit expériences ont été menées pour tenter de détecter des associations statistiquement significatives chez la population étudiée et pour comparer la méthode proposée à la correction de Bonferroni. Les expériences ont été choisies afin de comparer les deux méthodes dans plusieurs contextes. Nous avons étudié l'impact des facteurs suivants:

1. le nombre de SNPs étudiés simultanément;
2. le nombre de phénotypes étudiés simultanément;
3. la distribution des phénotypes;
4. la présence ou l'absence de liaison entre les SNPs et les phénotypes.

Les expériences ont aussi été choisies afin d'avoir une chance à priori élevée de détecter des associations significatives. Nous nous sommes basés sur la littérature pour choisir des régions et des phénotypes qui ont été rapportés comme étant liés, associés ou potentiellement associés. Ainsi, nous avons étudié 16 phénotypes hautement héritables sur les 22 autosomes (Hamet et al., 2005), l'obésité associée à l'hypertension sur le chromosome 1 (Pausova et al., 2005), les phénotypes anthropométriques sur le chromosome 1 (Hamet et al., 2005), la protéine C réactive sur le chromosome 10 et dans des gènes candidats (Broeckel et al., 2007), le syndrome métabolique et le gène FATP6 (Gimeno et al., 2003), le HDL et le chromosome 1 (Paré et al., 2007) et le gras corporel sur le chromosome 1 (Hamet et al., 2005).

8.2 Comparaison avec la correction de Bonferroni

Pour toutes les expériences, les valeurs p corrigées par la méthode proposée sont inférieures ou égales aux valeurs p corrigées par la méthode de Bonferroni.

La différence est parfois très élevée: pour l'expérience qui teste l'association entre les 22 autosomes et 16 phénotypes, la meilleure valeur p non corrigée, soit $9{,}1\times10^{-5}$ pour SUPRA, correspond à une valeur p corrigée de 0,03 selon notre méthode, comparativement à 1 selon Bonferroni. Dans ce cas précis, notre méthode est 30 fois moins sévère que la correction de Bonferroni.

Nous avons fait varier le nombre de marqueurs de cinq à 57 000, c'est-à-dire un rapport de quatre ordres de grandeur. Lorsqu'on examine chaque phénotype séparément, le rapport entre les valeurs p corrigées par Bonferroni et celles corrigées par notre méthode est le plus grand pour l'expérience qui contient le plus de marqueurs (voir tableau 7.4, colonne ratio_1). Lorsqu'on examine tous les phénotypes ensemble (colonne ratio_n), la plus grande différence est obtenue pour les expériences qui étudient 59 et 191 marqueurs. Il ne semble pas y avoir de relation simple entre le nombre de marqueurs et le rapport de puissance entre notre méthode et la correction de Bonferroni. Par contre, théoriquement, plus le LD est grand, plus notre méthode devrait se démarquer. L'arrivée récente des puces de génotypage de 1 million de SNPs pourrait permettre de vérifier cette théorie car le LD est beaucoup plus grand dans ces puces que dans la puce utilisée.

8.3 Association statistiquement significative

Notre méthode a permis de détecter une association statistiquement significative entre le SNP rs10494966 et le phénotype GRASP ($p_{corrigée} = 0{,}037$). Cette association a été jugée non significative par la correction de Bonferroni ($p_{corrigée} = 0{,}22$) et aurait donc été manquée si notre méthode n'avait pas été développée.

Cette association a été examinée afin de s'assurer qu'elle n'est pas le fruit d'une erreur de génotypage. Quatre vérifications ont été effectuées. Elles ont toutes fourni des réponses

satisfaisantes: le taux de génotypage de rs10494966 est de 99,6%; le test d'équilibre d'Hardy-Weinberg rapporte une valeur p de 0,79; la fréquence de l'allèle mineur est de 9,9%; aucune erreur n'a été détectée par la commande *allfreq* de *SOLAR*. Tous ces tests suggèrent que l'association n'est pas due à une erreur de génotypage.

Le SNP rs10494966 se trouve dans un intron du gène putatif Hs.368496 et est distant de 160 kb du gène PROX1 impliqué dans le cancer. Cette association est particulièrement intéressante du fait qu'elle se trouve à moins de 1 cM du pic de liaison de GRASP trouvé grâce aux microsatellites dans le même échantillon (Hamet et al., 2005). Le même signal est détecté par l'analyse de liaison par SNPs, ce qui donne confiance que le pic de liaison n'est pas dû à une erreur de génotypage.

8.4 Approche novatrice

Nous avons développé une approche de correction pour tests multiples dans les études d'association génomique familiale qui est très générale. Notre approche permet d'étudier simultanément un grand nombre de marqueurs, de phénotypes, de modèles génétiques et de sous-groupes d'individus. La valeur p corrigée tient compte de tous les tests qui sont effectués. La méthode peut être appliquée à n'importe quel test d'association familiale entre des génotypes et des phénotypes. Nous avons choisi le test FBAT comme test d'association de base, mais d'autres tests statistiques peuvent être intégrés à notre approche.

À notre connaissance, nous proposons la première méthode empirique de correction pour tests multiples dans les études d'association génomique familiale qui s'applique à des données aussi générales. Des méthodes empiriques simples, par exemple permuter librement les génotypes entre les individus ou encore permuter librement les phénotypes, ne seraient pas valides, soit parce que les lois mendéliennes seraient violées, soit parce que l'héritabilité ne serait pas conservée dans les simulations.

La méthode proposée est plus puissante que la correction de Bonferroni car elle tient compte des diverses corrélations qui existent entre les tests qui sont effectués, entre

autres la corrélation entre les marqueurs.

Notre approche respecte les trois conditions posées par Westfall et Young (1993) pour garantir la fiabilité du rééchantillonnage. Une validation expérimentale a d'ailleurs montré que la méthode contrôle adéquatement le FWER. La méthode proposée n'est donc pas libérale, contrairement à des méthodes plus simples inspirées des tests d'association cas-contrôle. Nous pouvons nous fier aux valeurs p corrigées qui sont rapportées, et les interpréter comme étant la probabilité d'observer un aussi bon résultat lorsqu'aucun SNP n'est associé aux phénotypes étudiés. Les valeurs p rapportées par les méthodes libérales ne possèdent pas cette propriété et sont donc de peu d'utilité.

Les méthodes de correction pour tests multiples actuelles basées sur les trios ou les familles ne permettent pas de traiter des familles étendues dans lesquelles plusieurs parents ne sont pas génotypés. À notre connaissance, seule la méthode proposée permet de traiter de telles familles. Cette propriété n'est pas superflue: une des familles de la population étudiée comprend à elle seule 96 individus dont 47 n'ont pas été génotypés.

Notre approche est pratique et se prête à une automatisation poussée: une expérience a été réalisée dans laquelle 1000 simulations ont été appliquées à un ensemble de plus de 57 000 marqueurs et à 16 phénotypes, soit près de un milliard de tests FBAT.

Afin de diminuer le temps écoulé entre le début et la fin d'une expérience, nous avons implanté un calcul distribué. Nous avons utilisé avec succès jusqu'à 17 processeurs à la fois. Puisqu'il existe un parallélisme de haut niveau que nous avons su exploiter, les opérations de communication sont beaucoup plus rapides que les opérations de calcul (en particulier la simulation d'haplotypes). Ainsi, l'utilisation de P processeurs réduit le temps de calcul par un facteur proche de P (accélération quasi-linéaire).

De plus, nous avons implanté une gestion des expériences afin, d'une part, conserver une trace des expériences effectuées et les répéter au besoin et, d'autre part, conserver les résultats des expériences et pouvoir facilement comparer les expériences entre elles ou comparer la méthode proposée à la correction de Bonferroni.

8.5 Simuler les génotypes ou les phénotypes?

Notre approche procède en simulant des génotypes qui respectent plusieurs propriétés des génotypes originaux (en particulier le LD et les fréquences alléliques) puis en testant l'association entre les génotypes simulés et les phénotypes originaux.

Une autre approche envisageable est de tester l'association entre des phénotypes simulés et les génotypes originaux. Il faut alors faire attention de conserver l'héritabilité: si ce n'est pas le cas, la correction peut ne pas être fiable, tel que nous l'avons décrit à la section 4.6. On pourrait conserver l'héritabilité en permutant les phénotypes au sein de la famille, c'est-à-dire en défendant l'échange entre deux individus qui ne proviennent pas de la même famille. Par contre, il pourrait y avoir peu de variabilité dans les petites familles.

8.6 Menaces à la validité

Nous avons affirmé que la méthode proposée contrôle adéquatement le FWER. Nous avons également déclaré une association statistiquement significative. Quelles erreurs d'expérimentation auraient pu causer ces résultats (validité interne) ou quels facteurs pourraient empêcher d'appliquer ces résultats à une population plus large (validité externe)?

8.6.1 Erreurs dans les logiciels

Les résultats obtenus dépendent du logiciel que nous avons développé et des logiciels tiers qui ont été intégrés. Des erreurs dans les logiciels pourraient produire les résultats rapportés. Nous avons tenté de diminuer cette probabilité de deux façons:

1. Des générateurs de nombre pseudo-aléatoires de bonne qualité ont été utilisés afin de diminuer l'erreur due à l'utilisation d'un générateur de nombres pseudo-aléatoires.
2. Le logiciel développé et deux logiciels intégrés ont été validés. Les logiciels se

sont comportés de manière attendue. La validation du logiciel développé est toutefois limitée: un seul jeu de données a été utilisé.

8.6.2 Violation des suppositions

Les principales suppositions de la méthode sont les suivantes:

1. Suffisamment de familles informatives sont fournies à *FBAT* pour que le théorème de la limite centrale s'applique (équation 3.3).
2. Les familles fournies à *FBAT* (hypothèse nulle H_{02}) sont faiblement apparentées.
3. Les individus fournis à *fastPHASE* sont faiblement apparentés.
4. Les informations utilisées pour guider la recherche sont statistiquement indépendantes des tests d'association effectués.

Concernant le nombre de familles informatives, nous avons exclu des analyses les marqueurs pour lesquels moins de 5 familles étaient informatives. Ce nombre est plus faible que les 10 familles recommandées par Laird (2006).

Pour que la statistique Z calculée à l'équation 3.3 soit valide, les scores des individus doivent être indépendants. Dans le cas de l'hypothèse nulle « pas d'association ni de liaison » (H_{01}), les scores des individus sont effectivement indépendants. Par contre, dans le cas de l'hypothèse nulle « pas d'association en présence de liaison » (H_{02}), les scores des individus sont corrélés à cause de la liaison. *FBAT* tient compte de cette corrélation au sein des familles en ajustant la variance de S dans l'équation 3.3. Par contre, *FBAT* ne fait aucun ajustement entre les familles. Or, nous savons que la plupart des familles de l'étude sont reliées à divers degrés. Cette corrélation viole la supposition du test H_{02}. Nous n'avons pas étudié l'impact de cette corrélation.

Le logiciel *fastPHASE*, utilisé pour phaser les génotypes des individus afin de calculer le LD entre les marqueurs, suppose que les individus sont faiblement apparentés, ce qui n'est pas le cas dans la présente étude. Les haplotypes inférés pourraient ne pas être les haplotypes les plus probables des individus. Nous n'avons pas étudié l'impact de cette

corrélation. Par contre, la validation de la méthode en son entier a montré qu'elle contrôle adéquatement le FWER, ce qui porte à croire que l'utilisation de *fastPHASE* n'a pas un impact négatif significatif.

8.6.3 Facteurs de confusion

Un facteur de confusion est une variable qui est associée autant à la variable indépendante (le génotype) qu'à la variable dépendante (le phénotype)[14]. Les principaux facteurs de confusion généralement étudiés (l'âge, le sexe, les habitudes de vie, etc.) ne sont pas à priori associés aux génotypes et ne sont donc pas considérés dans cette section. Par contre, les erreurs de génotypage peuvent être associées à la fois au phénotype et au génotype, par exemple lorsque les individus ayant des phénotypes similaires sont génotypés sur une même plaque. L'effet des erreurs de génotypage sur les résultats peuvent être énormes, comme l'ont démontré Knapp et Becker (2004) et Moskvina et al. (2006).

Nous avons mené quatre vérifications sur les génotypes du SNP associé: elles montrent que l'association rapportée n'est vraisemblablement pas due à une erreur de génotypage.

8.6.4 Généralisabilité

Les critères d'inclusion des sujets de l'étude, principalement la présence de deux membres hypertendus et dyslipidémiques au sein d'une fratrie, signifient que la population étudiée est en moins bonne santé que la population générale de la même région. L'association trouvée pourrait donc être moins forte dans la population générale.

Les individus de la population étudiée ont été recrutés à Chicoutimi et à Montréal et tous leurs parents sont nés au Saguenay-Lac-Saint-Jean. Un bagage génétique différent ou un environnement différent pourraient faire en sorte que cette association ne soit pas présente dans d'autres populations ou qu'elle ne soit pas aussi forte.

---

14 http://www.educ.necker.fr/cours/poly/biostatistique/Les_facteurs_de_confusion.htm

Chapitre 9: Conclusion

Nous avons inventé et implanté une méthode novatrice de correction pour tests multiples d'association génomique familiale entre des marqueurs génétiques et des phénotypes. La méthode contrôle fortement le taux d'erreur au sein de la famille de tests.

La méthode est basée sur le rééchantillonnage des génotypes. À chaque simulation, des génotypes sont simulés et testés pour l'association avec les phénotypes originaux. Les génotypes des fondateurs sont simulés en respectant les principales caractéristiques des génotypes originaux, soit les fréquences alléliques, les patrons de génotypes manquants et le déséquilibre de liaison. Puis, les génotypes des fondateurs sont transmis à leurs descendants par *gene-dropping*. Les phénotypes ne sont jamais modifiés.

La méthode développée s'applique aux études familiales dans lesquelles les familles peuvent être de taille arbitraire. Tous les patrons de génotypes manquants sont supportés, en particulier la situation dans laquelle les parents ne sont pas génotypés. Aucune variabilité phénotypique n'est nécessaire. La méthode supporte tout test d'association génomique familiale. À notre connaissance, nous proposons la première méthode empirique de correction pour tests multiples dans les études d'association génomique familiale qui s'applique à des données aussi générales.

Huit expériences ont été effectuées sur les données provenant de près de 900 sujets du Saguenay-Lac-Saint-Jean répartis en près de 120 familles. Ces données incluent un phénotypage étendu et, pour 468 d'entre eux, un génotypage de 58 000 SNPs par la puce Xba d'Affymetrix. Dans chaque expérience, la meilleure valeur p du test FBAT obtenue sur les données originales a été corrigée par la méthode proposée et par la correction de Bonferroni.

Dans toutes les expériences, les valeurs p corrigées par la méthode proposée sont inférieures ou égales aux valeurs p corrigées par Bonferroni. La différence est parfois très grande: dans une expérience portant sur les 22 autosomes, la meilleure valeur p

corrigée est de 0,03 selon notre méthode, comparativement à 1 selon la correction de Bonferroni, lorsqu'on considère chaque phénotype séparément.

Une association statistiquement significative après correction pour tests multiples a été détectée par notre méthode. Le SNP rs10494966 est associé au pourcentage de gras corporel calculé à partir des plis cutanés ($p_{corrigée} = 0{,}037$). En comparaison, la valeur p corrigée par Bonferroni est 0,22: cette association aurait été manquée si nous n'avions pas inventé et implanté la méthode proposée.

Le SNP rs10494966 est situé sur le chromosome 1 dans un intron du gène putatif Hs. 368496 et à 160 kb du gène connu le plus proche, PROX1. Quatre vérifications ont montré que l'association n'est pas due à une erreur de génotypage. Cette association démontre que notre approche peut détecter des associations statistiquement significatives même lorsque de nombreux tests d'association sont effectués.

Plusieurs validations ont été effectuées afin de s'assurer que la méthode proposée corrige correctement le taux d'erreur au sein de la famille de tests. Une validation faite sur $D = 4000$ répétitions de $N = 1000$ simulations d'un chromosome de 328 marqueurs a permis de constater que la méthode est fiable. En comparaison, la correction de Bonferroni corrige trop fortement les valeurs p et s'avère conservatrice.

Plusieurs avenues restent à explorer: une méthode de simulations alternative est envisageable, des validations plus poussées permettraient d'augmenter notre confiance dans la méthode et dans le logiciel et un autre test d'association de base pourrait être plus puissant.

Premièrement, nous avons rééchantillonné les génotypes tout en gardant les phénotypes originaux. Une alternative est de garder les génotypes originaux et de rééchantillonner les phénotypes. Il faut toutefois s'assurer que les phénotypes simulés conservent l'héritabilité des phénotypes originaux.

Deuxièmement, une validation plus poussée de la méthode serait la bienvenue et nous donnerait davantage confiance dans ses résultats. En particulier, la validation effectuée a

été faite sur une seule structure généalogique, un seul groupe de génotypes, une seule simulation de phénotypes et une seule taille d'échantillon. Il serait important de faire varier ces paramètres et s'assurer que la méthode corrige adéquatement le taux d'erreur au sein de la famille de tests dans une large gamme de conditions expérimentales.

Finalement, le test FBAT appliqué aux données utilisées dans les expériences semble manquer de puissance. Un autre test d'association génomique familiale pourrait être plus puissant, donc produire plus de résultats d'association positifs. Il serait intéressant de trouver dans la littérature un tel test, ou même d'en développer un, puis d'incorporer ce test à notre méthode.

Chapitre 10: Références

Abecasis, G. R., Cherny, S. S., Cookson, W. O., & Cardon, L. R. (2002). Merlin - rapid analysis of dense genetic maps using sparse gene flow trees. *Nature Genetics, 30*(1), 97-101.

Abney, M., Ober, C., & McPeek, M. S. (2002). Quantitative-trait homozygosity and association mapping and empirical genomewide significance in large, complex pedigrees: fasting serum-insulin level in the Hutterites. *American Journal of Human Genetics, 70*(4), 920-934.

Affymetrix. (2006). *README: 500K Copy Number Sample Data Sets*. Santa Clara, California: Affymetrix.

Almasy, L., & Blangero, J. (1998). Multipoint quantitative-trait linkage analysis in general pedigrees. *American Journal of Human Genetics, 62*(5), 1198-1211.

Anderson, M. J. (2001). Permutation tests for univariate or multivariate analysis of variance and regression. *Canadian Journal of Fisheries and Aquatic Sciences, 58*(3), 626-639.

Barrett, J. C., Fry, B., Maller, J., & Daly, M. J. (2005). Haploview: analysis and visualization of LD and haplotype maps. *Bioinformatics, 21*(2), 263-265.

Barroso, I., Luan, J., Middelberg, R. P. S., Harding, A.-H., Franks, P. W., Jakes, R. W., Clayton, D., Schafer, A. J., O'Rahilly, S., & Wareham, N. J. (2003). Candidate gene association study in type 2 diabetes indicates a role for genes involved in β-cell function as well as insulin action. *PLoS Biology, 1*(1), e20.

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society, 57*(1), 289-300.

Benjamini, Y., & Yekutieli, D. (2001). The control of the false discovery rate in multiple testing under dependency. *The Annals of Statistics, 29*(4), 1165-1188.

Berger, J. O., & Sellke, T. (1987). Testing a point null hypothesis: the irreconcilability of p values and evidence. *Journal of the American Statistical Association, 82*(397), 112-122.

Bouchard, G. (2006). *Rapport annuel du projet BALSAC 2005-2006.* Saguenay: Université du Québec à Chicoutimi.

Bowdler, H., Martin, R. S., Reinsch, C., & Wilkinson, J. H. (1968). The QR and QL algorithms for symmetric matrices. *Numerische Mathematik, 11*, 293-306.

Broeckel, U., Hengstenberg, C., Mayer, B., Maresso, K., Gaudet, D., Seda, O., Tremblay, J., Holmer, S., Erdmann, J., Glöckner, C., Harrison, M., Martin, L. J., Williams, J. T., Schmitz, G., Riegger, G. A., Jacob, H. J., Hamet, P., & Schunkert, H. (2007). A locus on chromosome 10 influences C-reactive protein levels in two independent populations. *Human Genetics, 122*(1), 95-102.

Churchill, G. A., & Doerge, R. W. (1994). Empirical threshold values for quantitative trait mapping. *Genetics 138*(3), 963-971.

Curran-Everett, D. (2000). Multiple comparisons: philosophies and illustrations. *American Journal of Physiology - Regulatory, Integrative and Comparative Physiology, 279*(1), R1-R8.

Danesh, J., Wheeler, J. G., Hirschfield, G. M., Eda, S., Eiriksdottir, G., Rumley, A., Lowe, G. D. O., Pepys, M. B., & Gudnason, V. (2004). C-reactive protein and other circulating markers of inflammation in the prediction of coronary heart disease. *The New England Journal of Medicine, 350*(14), 1387-1397.

Devlin, B., & Risch, N. (1995). A comparison of linkage disequilibrium measures for fine-scale mapping. *Genomics, 29*(2), 311-322.

Diabetes Genetics Initiative of Broad Institute of Harvard and MIT, Lund University, and Novartis Institutes of BioMedical Research, Saxena, R., Voight, B. F., Lyssenko, V., Burtt, N. P., de Bakker, P. I., Chen, H., Roix, J. J., Kathiresan, S., Hirschhorn, J. N., Daly, M. J., Hughes, T. E., Groop, L., Altshuler, D., Almgren, P., Florez, J. C., Meyer, J., Ardlie, K., Bengtsson Boström, K., Isomaa, B., Lettre, G., Lindblad, U., Lyon, H. N., Melander, O., Newton-Cheh, C., Nilsson, P., Orho-Melander, M., Råstam, L., Speliotes, E. K., Taskinen, M. R., Tuomi, T., Guiducci, C., Berglund, A., Carlson, J., Gianniny, L., Hackett, R., Hall, L., Holmkvist, J., Laurila, E., Sjögren, M., Sterner, M., Surti, A., Svensson, M., Svensson, M., Tewhey, R., Blumenstiel, B., Parkin, M., Defelice, M., Barry, R., Brodeur, W., Camarata, J., Chia, N., Fava, M., Gibbons, J., Handsaker, B., Healy, C., Nguyen, K., Gates, C., Sougnez, C., Gage, D., Nizzari, M., Gabriel, S. B., Chirn, G. W., Ma, Q., Parikh, H., Richardson, D., Ricke, D., & Purcell, S. (2007). Genome-wide association analysis identifies loci for type 2 diabetes and triglyceride levels. *Science, 316*(5829), 1331-1336.

Dudbridge, F., & Koeleman, B. P. C. (2003). Rank truncated product of p-values, with application to genomewide association scans. *Genetic Epidemiology, 25*(4), 360-366.

Dudbridge, F., Gusnanto, A., & Koeleman, B. P. C. (2006). Detecting multiple associations in genomewide studies. *Human Genomics, 2*(5), 310-317.

Elston, R. C. (2001). Reporting of linkage results. *American Journal of Human Genetics, 69*(5), 1149-1150.

Efron, B. (2004). Large-Scale simultaneous hypothesis testing: the choice of a null hypothesis. *Journal of the American Statistical Association, 99*(465), 96-104.

Entacher, K., Uhl, A., & Wegenkittl, S. (1998). Linear and inversive pseudorandom numbers for parallel and distributed simulation. *Proceedings of the twelfth workshop on parallel and distributed computing*, Banff, Alberta, Canada, 90-97.

Franke, D., Kleensang, A., & Ziegler, A. (2006). SIBSIM - quantitative phenotype simulation in extended pedigrees. *GMS Medizinische Informatik, Biometrie und Epidemiologie, 2*(1), Doc02.

Freidlin, B., Zheng, G., Li, Z., & Gastwirth, J. L. (2002). Trend tests for case-control studies of genetic markers: power, sample size and robustness. *Human Heredity, 53*(3), 146-152.

Ge, Y., Dudoit, S., & Speed, T. P. (2003). *Resampling-based multiple testing for microarray data analysis* (TR 633). Berkeley: University of California at Berkeley.

Gigerenzer, G. (2004). Mindless statistics. *The Journal of Socio-Economics, 33*(5), 587-606.

Gimeno, R. E., Ortegon, A. M., Patel, S., Punreddy, S., Ge, P., Sun, Y., Lodish, H. F., & Stahl, A. (2003). Characterization of a heart-specific fatty acid transport protein. *Journal of Biological Chemistry, 278*(18), 16039-16044.

Goeman, J. J., van de Geer, S. A., de Kort, F., & van Houwelingen, H. C. (2004). A global test for groups of genes: testing association with a clinical outcome. *Bioinformatics, 20*(1), 93-99.

Hall, P., & Wilson, S. R. (1991). Two guidelines for bootstrap hypothesis testing. *Biometrics, 47*(2), 757-762.

Hauser, E. R., Watanabe, R. M., Duren, W. L., Bass, M. P., Langefeld, C. D., & Boehnke, M. (2004). Ordered subset analysis in genetic linkage mapping of complex traits. *Genetic Epidemiology, 27*(1), 53-63.

Hamet, P., Merlo, E., Seda, O., Broeckel, U., Tremblay, J., Kaldunski, M., Gaudet, D., Bouchard, G., Deslauriers, B., Gagnon, F., Antoniol, G., Pausova, Z., Labuda, M., Jomphe, M., Gossard, F., Tremblay, G., Kirova, R., Tonellato, P., Orlov, S. N., Pintos, J., Platlo, J., Hudson, T. J., Rioux, J. D., Kotchen, T. A., & Cowley, A. W. Jr. (2005). Quantitative founder-effect analysis of French Canadian families identifies specific loci contributing to metabolic phenotypes of hypertension. *American Journal of Human Genetics, 76*(5), 815-832.

Heather, L. C., Cole, M. A., Lygate, C. A., Evans, R. D., Stuckey, D. J., Murray, A. J., Neubauer, S., & Clarke, K. (2006). Fatty acid transporter levels and palmitate oxidation rate correlate with ejection fraction in the infarcted rat heart. *Cardiovascular Research, 72*(3), 430-437.

Hellekalek, P. (1998). Don't trust parallel Monte Carlo! *Proceedings of the twelfth workshop on parallel and distributed computing*, Banff, Alberta, Canada, 82-89.

Horvath, S., Xu X., Lake, S. L., Silverman, E. K., Weiss, S. T., & Laird, N. M. (2004). Family-based tests for associating haplotypes with general phenotype data: application to asthma genetics. *Genetic Epidemiology, 26*(1), 61-69.

Huang, Y., Xu, H., Calian, V., & Hsu, J. C. (2006). To permute or not to permute. *Bioinformatics, 22*(18), 2244-2248.

International Human Genome Sequencing Consortium. (2004). Finishing the euchromatic sequence of the human genome. *Nature, 431*(7011), 931-945.

Ioannidis, J. P. A. (2005). Why most published findings are false. *PLoS Medicine, 8*(2), 696-701.

Ishwaran, H., & Rao, J. S. (2003). Detecting differentially expressed genes in microarrays using bayesian model classification. *Journal of the American Statistical Association, 98*(462), 438-455.

Jung J., Weeks D. E., & Feingold E. (2006). Gene-dropping vs. empirical variance estimation for allele-sharing linkage statistics. *Genetic Epidemiology, 30*(8), 652-665.

Kimmel, G., & Shamir, R. (2006). A fast method for computing high-significance disease association in large population-based studies. *American Journal of Human Genetics, 79*(3), 481-492.

Knapp, M., & Becker, T. (2004). Impact of genotyping errors on type I error rate of the haplotype-sharing transmission/disequilibrium test (HS-TDT). *American Journal of Human Genetics, 74*(3), 589-591.

Kruglyak, L., & Daly, M. J. (1998). Linkage thresholds for two-stage genome scans. *American Journal of Human Genetics, 62*(4), 994-996.

Laird, N. M. (2006). *Family-based association tests and the FBAT-toolkit: user's manual.* Boston: Harvard.

Lander, E., & Kruglyak, L. (1995). Genetic dissection of complex traits: guidelines for interpreting and reporting linkage results. *Nature Genetics, 11*(3), 241-247.

Lange, K., Cantor, R., Horvath, S., Papp, J. C., Sabatti, C., Sinsheimer, J. S., & Sobel E. (2005). *Mendel 6.0 documentation*. Los Angeles: University of California at Los Angeles.

Lazzeroni, L. C., & Lange, K. (1998). A conditional inference framework for extending the transmission/disequilibrium test. *Human Heredity*, 48(2), 67-81.

Lin, D. Y., & Zou, F. (2004). Assessing genomewide statistical significance in linkage studies. *Genetic Epidemiology, 27*(3), 202-214.

Lin, D. Y. (2005). An efficient Monte Carlo approach to assessing statistical significance in genomic studies. *Bioinformatics, 21*(6), 781-787.

Lin, D. Y. (2006). Evaluating statistical significance in two-stage genomewide association studies. *American Journal of Human Genetics, 78*(3), 505-509.

Lodish, H., Baltimore, D., Berk, A., Zipursky, S. L., Matsudaira, P., & Darnell, J. (1997). *Biologie moléculaire de la cellule* (3e éd.). Paris: De Boeck Université.

MacCluer, J. W., VandeBerg, J. L., Read, B., & Ryder, O. A. (1986). Pedigree analysis by computer simulation. *Zoo Biology, 5*(2), 147-160.

Martin, R. S., Reinsch, C., & Wilkinson, J. H. (1968). Householder's tridiagonalization of a symmetric matrix. *Numerische Mathematik, 11*(3), 181-195.

Maller, J., George, S., Purcell, S., Fagerness, J., Altshuler, D., Daly, M. J., & Seddon, J. M. (2006). Common variation in three genes, including a noncoding variant in CFH, strongly influences risk of age-related macular degeneration. *Nature Genetics, 38*(9), 1055-1059.

Matsumoto, M., & Nishimura, T. (1998). Mersenne-Twister: a 623-dimensionally equidistributed uniform pseudorandom number generator. *ACM Transactions on Modeling and Computer Simulation, 8*(1), 3-30.

Mattick, J. S., & Igor, V. M. (2006). Non-coding RNA. *Human Molecular Genetics, 15*(1), R17-R29.

McIntyre, L. M., Martin, E. R., Simonsen, K. L., & Kaplan, N. L. (2000). Circumventing multiple testing: a multilocus Monte Carlo approach to testing for association. *Genetic Epidemiology, 19*(1), 18-29.

Montana, G. (2005). HapSim: a simulation tool for generating haplotype data with pre-specified allele frequencies and LD coefficients. *Bioinformatics, 21*(23), 4309-4311.

Moskvina, V., Craddock, N., Holmans, P., Owen, M. J., & O'Donovan, M. C. (2006). Effects of differential genotyping error rate on the type I error probability of case-control studies. *Human Heredity, 61*(1), 55-64.

National Cholesterol Education Program (NCEP). (2002). *Detection, evaluation, and treatment of high blood cholesterol in adults (ATP-III)*. Bethesda, Maryland: National Institutes of Health.

Newman, D. L., Abney, M., McPeek, M. S., Ober, C., & Cox, N. J. (2001). The importance of genealogy in determining genetic associations with complex traits. *American Journal of Human Genetics, 69*(5), 1146-1148.

Nyholt, D. R. (2004). A simple correction for multiple testing for single-nucleotide polymorphisms in linkage disequilibrium with each other. *American Journal of Human Genetics, 74*(4), 765-769.

Oppert, J.-M. (2003). Obésités: quelles mesures pour les « phénotypes » à risque cardiovasculaire? *Sang Trombose Vaisseaux, 15*(9-10), 551-556.

Ostle, B., Turner, K. V. Jr., Hicks, C. R., & McElrath, G. W. (1996). *Engineering statistics: the industrial experience.* Belmont, California: Duxbury Press.

Ott, J. (1991). *Analysis of human genetic linkage* (2e éd.). Baltimore, Maryland: The Johns Hopkins University Press.

Paré, G., Serre, D., Brisson, D., Anand, S. S., Montpetit, A., Tremblay, G., Engert, J. C., Hudson, T. J., & Gaudet, D. (2007). Genetic analysis of 103 candidate genes for coronary artery disease and associated phenotypes in a founder population reveals a new association between endothelin-1 and high-density lipoprotein cholesterol. *American Journal of Human Genetics, 80*(4), 673-682.

Pausova, Z., Deslauriers, B., Gaudet, D., Tremblay, J., Kotchen, T. A., Larochelle, P., Cowley, A. W., & Hamet, P. (2000). Role of tumor necrosis factor-a gene locus in obesity and obesity-associated hypertension in French Canadians. *Hypertension, 36*(1), 14-19.

Pausova, Z., Gaudet, D., Gossard, F., Bernard, M., Kaldunski, M. L., Jomphe, M., Tremblay, J., Hudson, T. J., Bouchard, G., Kotchen, T. A., Cowley, A. W., & Hamet, P. (2005). Genome-wide scan for linkage to obesity-associated hypertension in French Canadians. *Hypertension, 46*(6), 1280-1285.

Pounds, S., & Morris, S. W. (2003). Estimating the occurrence of false positives and false negatives in microarray studies by approximating and partitioning the empirical distribution of p-values. *Bioinformatics, 19*(10), 1236-1242.

Pressman, R. S. (1997). *Software engineering: a practitioner's approach* (4e éd.). New York, New York: McGraw Hill.

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A. R., Bender, D., Maller, J., Sklar, P., de Bakker, P. I. W., Daly, M. J., & Sham, P. C. (2007). PLINK: a toolset for whole-genome association and population-based linkage analysis. *American Journal of Human Genetics, 81*(3), 559-575.

R Development Core Team. (2006). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing.

Redon, R., Ishikawa, S., Fitch, K. R., Feuk, L., Perry, G. H., Andrews, T. D., Fiegler, H., Shapero, M. H., Carson, A. R., Chen, W., Cho, E. K., Dallaire, S., Freeman, J. L., González, J. R., Gratacòs, M., Huang, J., Kalaitzopoulos, D., Komura, D., MacDonald, J. R., Marshall, C. R., Mei, R., Montgomery, L., Nishimura, K., Okamura, K., Shen, F., Somerville, M. J., Tchinda, J., Valsesia, A., Woodwark, C., Yang, F., Zhang, J., Zerjal, T., Zhang, J., Armengol, L., Conrad, D. F., Estivill, X., Tyler-Smith, C., Carter, N. P., Aburatani, H., Lee, C., Jones, K. W., Scherer, S. W., & Hurles, M. E. (2006). Global variation in copy number in the human genome. *Nature, 444*(7118), 444-454.

Ridker, P. M., Hennekens, C. H., Buring, J. E., & Rifai, N. (2000). C-reactive protein and other markers of inflammation in the prediction of cardiovascular disease in women. *The New England Journal of Medicine, 342*(12), 836-843.

Roeder, K., Bacanu, S.-A., Wasserman, L., & Devlin, B. (2006). Using linkage genome scans to improve power of association in genome scans. *American Journal of Human Genetics, 78*(2), 243-252.

Sabatti, C., Service, S., & Freimer, N. (2003). False discovery rate in linkage and association genome screens for complex disorders. *Genetics, 164*(2), 829-833.

Salyakina, D., Seaman, S. R., Browning, B. L., Dudbridge, F., & Muller-Myhsok, B. (2005). Evaluation of Nyholt's procedure for multiple testing correction. *Human Heredity, 60*(1), 19-25.

Sawcer, S., Jones, H. B., Judge, D., Visser, F., Compston, A., Goodfellow, P. N., & Clayton, D. (1997). Empirical genomewide significance levels established by whole genome simulations. *Genetic Epidemiology, 14*(3), 223-229.

S.A.G.E. (2006). Statistical Analysis for Genetic Epidemiology, Release 5.3: http://genepi.cwru.edu.

Scheet, P., & Stephens, M. (2006). A fast and flexible statistical model for large-scale population genotype data: applications to inferring missing genotypes and haplotypic phase. *American Journal of Human Genetics, 78*(4), 629-644.

Seda, O., Tremblay, J., Gaudet, D., Brunelle, P.-L., Gurau, A., Merlo, E., Pilote, L., Orlov, S. N., Boulva, F., Petrovich, M., Kotchen, T. A., Cowley, A. W. Jr., & Hamet, P. (2008). Systematic, genome-wide, sex-specific linkage of cardiovascular traits in French Canadians. *Hypertension, 51*(4), 1156-1162.

Serena, D., Passarino, G., Rose, G., Altomare, K., Bellizzi, D., Mari, V., Feraco, E., Franceschi, C. & De Benedictis, G. (2004). Association of the mitochondrial DNA haplogroup J with longevity is population specific. *European Journal of Human Genetics, 12*(12), 1080-1082.

Sladek, R., Rocheleau, G., Rung, J., Dina, C., Shen, L., Serre, D., Boutin, P., Vincent, D., Belisle, A., Hadjadj, S., Balkau, B., Heude, B., Charpentier, G., Hudson, T. J., Montpetit, A., Pshezhetsky, A. V., Prentki, M., Posner, B. I., Balding, D. J., Meyre, D., Polychronakos, C., & Froguel, P. (2007). A genome-wide association study identifies novel risk loci for type 2 diabetes. *Nature, 445*(7130), 881-885.

Storey, J. D. (2002). A direct approach to false discovery rates. *Journal of the Royal Statistical Society, 64*, 479-498.

Terwilliger, J. D., & Ott, J. (1992). A multisample bootstrap approach to the estimation of maximized-over-models lod score distributions. *Cytogenetics and Cell Genetics, 59*(2-3), 142-144.

Tremblay, J. (2007). Genetic determinants of C-reactive protein levels in metabolic syndrome: a role for the adrenergic system? *Journal of Hypertension, 25*(2), 281-283.

Tremblay, M., Arsenault, J., & Heyer, E. (2003). Les probabilités de transmission des gènes fondateurs dans cinq populations régionales du Québec. *Population, 58*(3), 403-423.

Van Steen, K., McQueen, M. B., Herbert, A., Raby, B., Lyon, H., DeMeo, D. L., Murphy, A., Su, J., Datta, S., Rosenow, C., Christman, M., Silverman, E. K., Laird, N. M., Weiss, S. T., & Lange, C. (2005). Genomic screening and replication using the same data set in family-based association testing. *Nature Genetics, 37*(7), 683-691.

Wacholder, S., Chanock, S., Garcia-Closas, M., El Ghormli, L., & Rothman, N. (2004). Assessing the probability that a positive report is false: an approach for molecular epidemiology studies. *Journal of the National Cancer Institute, 96*(6), 434-442.

Wellcome Trust Case Control Consortium. (2007). Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature, 447*(7145), 661-678.

Westfall, P. H., & Young, S. S. (1993). *Resampling-based multiple testing: examples and methods for p-value adjustment.* New York, New York: John Wiley & Sons.

Wu, X., & Naiman, D. Q. (2005). P-value simulation for affected sib pair multiple testing. *Human Heredity, 59*(4), 190-200.

Zaykin, D. V., Zhivotovsky, L. A., Westfall, P. H., & Weir, B. S. (2002). Truncated product method for combining p-values. *Genetic Epidemiology, 22*(2), 170-185.

Zhao, J. H., Curtis, D., & Sham, P. K. (2000). Model-free analysis and permutation tests for allelic associations. *Human Heredity, 50*(2), 133-139.